ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ
UNIVERSITÉ DE LA MÉDITERRANÉE AIX-MARSEILLE II
FACULTÉ DES SCIENCES DE LUMINY

THÈSE

pour obtenir le titre de

Docteur en Sciences

Spécialité : BIOINFORMATIQUE ET GÉNOMIQUE

Présentée et soutenue par

Aurélie BERGON

Développement d'outils et méthodes bioinformatiques pour l'étude de l'expression des gènes et de leur régulation : Application aux pathologies

soutenue le 6 février 2012

Jury :

Rapporteurs :	Pr. Gianluca BONTEMPI	-	Université Libre de Bruxelles
	Dr. Frédéric GUYON	-	Inserm UMR_S 973 (Paris)
Examinateurs :	Dr. Max CHAFFANET	-	CRCM Inserm UMR891 (Marseille)
	Dr. Salvatore SPICUGLIA	-	Inserm UMR_S 928 (Marseille)
Directeurs :	Dr. Jean IMBERT	-	Inserm UMR_S 928 (Marseille)
	Dr. Denis PUTHIER	-	Inserm UMR_S 928 (Marseille)
Président :	Pr. Franck GALLAND	-	CIML (Marseille)

Remerciements

Tout d'abord, je tiens à remercier les membres de mon jury d'avoir accepté de lire et de juger mon travail de Doctorat, malgré le court délai que je leur ai finalement laissé. Je voudrais ensuite exprimer ma gratitude à la directrice du laboratoire Inserm UMR_S 928 TAGC, le Docteur Catherine Nguyen, pour m'avoir accueillie depuis 5 ans. Je remercie également mes directeurs de thèse, les docteurs Jean Imbert et Denis Puthier, pour m'avoir permis de prendre part à des projets de recherche fort intéressants, qui m'ont beaucoup appris.

Un immense merci à mes amis de la plateforme TGML, pour votre amitié et pour tous les moments de détente et les fous rires inoubliables partagés ensemble. A Fabrice et FX, mes collègues bioinformaticiens ! ! ! ah, on en passe du temps sur les données de séquençage, et ce n'est que le début, à très bientôt pour poursuivre le développement. Et en particulier à Fabrice pour son aide en programmation lors de ma thèse, on va bientôt pouvoir reprendre les soirées pizzas... A Hélène, pour toute son aide et sa patience afin de m'expliquer et surtout de m'aider à venir à bout des expériences, notamment du ChIP-seq et pour toute la culture cellulaire qu'elle a réalisée pour moi. A Valérie, pour sa gentillesse et les bons gâteaux au chocolat, un antidépresseur naturel qui m'a été très utile, lors de cette thèse.

A Sophie et Véro, notre équipe de choc de gestion, merci les filles pour les pauses cafés, qui m'ont été d'un grand réconfort. Merci en particulier à toi Sophie, pour toutes les relectures que tu as faites malgré le fait que la bioinformatique ne soit pas ta tasse de thèse... euh je veux dire de thé ! ! !

A Clairette, merci de m'avoir encouragée en me montrant qu'on pouvait vaincre toutes les difficultés... Merci pour m'avoir encouragée à aller prendre des cours de salsa avec toi, sans quoi je n'aurais jamais rencontré l'amour de ma vie, Christophe. J'en profite pour remercier tous mes amis de la salsa, de la bachata et de la kizomba... à très bientôt sur les pistes de danse, espérons que d'ici là je n'aie pas tout oublié.

J'adresse mes remerciements à tous mes collaborateurs auprès desquels j'ai beaucoup appris au cours de ces 4 années. En particulier, je souhaite vivement remercier le Dr El Chérif Ibrahim pour ses conseils, son aide précieuse et les critiques qu'il a pu apporter lors de la laborieuse rédaction de ce manuscrit.

A mes amies de Prédiguard Angela et Florence, et à tous mes amis du laboratoire présents et passés : Jacky, Brigitte, Alex, Laura, Mimz, Luca, Sève, Nath, Cyrille, et tous les autres ; un grand merci pour votre soutien et tous les bons moments de rigolade et pour l'ambiance de travail unique du TAGC. A Jacques et Sam, qui malgré leur emploi du temps très chargé m'ont accordé du temps lors de nombreuses discussions constructives et pour leurs conseils.

A mes amis Martine et Jean-Louis pour leur bons sandwichs qui m'ont nourri pendant ces années de thèse à des horaires souvent hors normes, merci pour votre amitié.

Une pensée à ma grande famille et à celle de Christophe, en particulier à mes parents et

à mon frère Olivier. Vous avez toujours été là pour moi, même toi Olivier parti vivre loin de nous en Chine ! ! ! Merci pour votre soutien et le réconfort que vous m'avez apporté pendant les moments de doute et de stress de ces dernières années, et pour m'avoir poussée toujours en avant.

Enfin, un immense merci à mon chéri Christophe, qui a subi ces 9 derniers mois la naissance de cette thèse, pour tout le soutien que tu m'as toujours donné. On a connu le plus dur mon coeur, il nous reste le meilleur à vivre.

A mes parents pour leur immense soutien,
et en mémoire de ma grand mère Monique
qui m'a toujours poussée à me surpasser

Table des matières

Liste des figures

Liste des tables

Liste des abréviations

Les abréviations indiquées ci-dessous sont en anglais, car ce sont celles communément admise par la communauté scientifique.

AFE	Agilent Feature Extraction software
ANOVA	ANalysis Of VAriance
ChIP	Chromatin ImmunoPrecipitation
FDR	False Discovery Rate
Gb	Gigabytes
GEO	Gene Expression Omnibus
HTS	High Thoughput Sequencing
ICS	Instrument Controler Software
LOWESS	LOcaly WEighted Scatterplot Smoothing
nt	Nucleotide
PCR	Polymerase Chain Reaction
PET	Paired-End Tag
RNA	RiboNucleic Acid (ou ARN en français)
SAM	Significant Analysis of Microarrays
SETS	SOLiD Experimental Tracking Software
SNP	Small Nucleotide Polymorphism
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
Tb	Terabytes
TS	TranscriptomeBrowser's Transcriptional Signature
TSS	Transcription Start Site

Résumé de la thèse

Suite à des études en biologie effectuées à l'IUT Génie Biologie (Analyses Biologiques et Biochimiques) puis à l'Université de Toulon et du Var, j'ai débuté ma formation en bioinformatique en 2006 en intégrant le master 1 mention **Bioinformatique, Biochimie Structurale et Génomique (BBSG)** à la Faculté des Sciences de Luminy (Université de la Méditerranée, AixMarseille II). Lors de ce cursus j'ai réalisé deux stages en bioinformatique au sein de l'unité mixte Inserm/Université de la Méditerranée UMR_S 928 intitulé Technologies Avancées pour le Génome et la Clinique (TAGC) sous la direction du Dr Denis Puthier puis en co-direction avec le Dr Jean Imbert, qui a rejoint le TAGC en juillet 2007.

Le **laboratoire TAGC** mène des projets de recherche dans les domaines de la génomique et de la bioinformatique avec pour la plupart une application médicale, et accueille une plateforme de Transcriptomique et de Génomique labélisée IBiSA basée sur la technologie des puces à ADN et du séquençage à très haut débit, intitulée Transcriptome Génomique Marseille Luminy (TGML). Ces projets combinent des analyses portant sur la régulation transcriptionnelle notamment dans le cancer du sein, les lymphomes, les glioblastomes, ou encore dans le sepsis. Ce laboratoire effectue également des recherches fondamentales sur le contrôle du cycle cellulaire, de la différenciation et de l'activation des lymphocytes T chez les mammifères. Ces projets combinent des analyses portant sur la transcription et la régulation de l'expression des gènes, et impliquent le développement d'outils d'analyse et d'approches bioinformatiques.

Les stages pré-doctoraux m'ont permis d'acquérir une expérience dans (1) l'analyse de données de puces à ADN avec le développement d'une librairie R permettant le contrôle qualité et la normalisation des puces sur lame de verre Agilent™, et (2) la méta-analyse de données de puces à ADN provenant de Gene Expression Omnibus (GEO) avec ma participation au projet TranscriptomeBrowser initié en 2007 par le Docteur Denis Puthier. Ce dernier projet m'a permis d'étudier la co-expression de gènes et leur régulation dans le cadre de l'activation et de la différenciation des lymphocytes T. Suite à l'obtention en Juillet 2008 d'**une bourse d'allocataire de recherche MRT**, j'ai souhaité poursuive mes travaux de recherche au TAGC. J'ai ainsi effectué ma thèse en co-direction avec les docteurs Jean Imbert et Denis Puthier.

Dans la continuité de mes travaux de master, ma thèse a eu pour objet le développement d'outils et de méthodes bioinformatiques permettant la caractérisation des mécanismes impliqués dans diverses maladies par une approche transcriptomique pan-génomique, mais également épigénétique et par l'étude de la régulation transcriptionnelle de l'expression des gènes.

J'ai évalué **la qualité des données de puces à ADN Agilent™, je les ai normalisées et analysées dans le cadre de collaborations** avec d'autres équipes de recherche. J'ai ainsi travaillé avec le Dr. Patricia Paris de l'Institut de Médecine Tropicale du Service de Santé des Armées (IMTSSA, Marseille) sur la dengue, infection virale, et avec le Dr. El Chérif Ibrahim du NICN CNRS UMR 6184 (Faculté de Médecine Nord, Marseille) sur l'étude d'une maladie orpheline et neurodégénérative, la dysautonomie familiale. **Ces collaborations ont fait l'objet de trois publications : deux dans PLoS ONE et une dans Human Mutation** (voir Chapitres

2 et 3).

Le **projet TranscriptomeBrowser, publié en décembre 2008 dans PLoS ONE**, a été poursuivi. J'ai ainsi développé de nouvelles fonctionnalités, mais également restructuré, optimisé et mis à jour notre base de données. **Ces travaux font l'objet d'un article accepté dans BMC Bioinformatics et d'un autre en préparation** (voir Chapitre 4).

En avril 2009, l'évolution technologique marquée par l'arrivée d'un séquenceur à très haut débit SOLiD^{TM}v3 sur la plateforme TGML m'a permis de **développer un pipeline d'analyse pour des expériences d'ImmunoPrécipitation de la Chromatine (ChIP-seq)** (voir Chapitre 5). Ce développement bioinformatique m'a ensuite permis de collaborer à l'analyse d'expériences ciblant de facteurs de transcription impliqués dans le cancer du sein ou les glioblastomes, respectivement avec l'équipe du Dr Daniel Birnbaum du Centre de Recherche en Cancérologie de Marseille (CRCM) et le Dr Thierry Virolle de l'unité Inserm U898 « *stem cells, development and cancer* » de Nice.

Enfin, le pipeline et les scripts développés ont également été utilisés dans le cadre de l'analyse de données de positionnement des nucléosomes par une approche développée par le Dr Salvatore Spicuglia de l'équipe du Dr Pierre Ferrier au CIML en collaboration avec notre laboratoire. **Cette technique nommée Mnase-Cap fait l'objet d'un article en préparation.**

Chapitre 1

Introduction générale

Sommaire

1.1 Étude des pathologies

La pathologie est une partie de la médecine qui a pour objet l'étude des maladies et notamment leurs causes, leurs mécanismes, leurs développements et leurs symptômes. Un abus de langage relativement récent et populaire consiste à faire du mot « pathologie » un synonyme du mot « maladie », en l'utilisant ainsi pour toute altération pathologique d'un mécanisme ou d'un processus biologique.

La plupart des maladies sont multifactorielles, c'est-à-dire quelles possèdent plusieurs altérations ou causes. Leur occurrence dépend de l'environnement (dans le cas de l'infection par exemple), du vécu de l'individu, mais aussi des prédispositions que lui confère son patrimoine génétique pour des maladies héréditaires. Dans ce cas, les facteurs génétiques ne font que prédisposer un individu au déclenchement et au développement particulier d'une pathologie, mais n'impliquent en aucun cas son apparition.

Ces maladies complexes sont diverses, telles que : obésité, diabète, asthme, cancer, maladies auto-immunes, maladies neurodégénératives, etc. Des examens médicaux, comme des prises

de sang et autres examens cliniques, permettent de les détecter et de suivre leur évolution, tant au niveau d'un tissu ou organe donné que de tout l'organisme. En effet, toutes les maladies ont des évolutions et des pronostics différents.

Les travaux de recherche sur ces maladies visent à la compréhension de leurs perturbations. Ainsi, l'étude à grande échelle de l'expression des gènes permet de déterminer des gènes différentiellement exprimés pouvant expliquer les perturbations observées. Ces travaux conduisent, entre autres, à des analyses fonctionnelles au niveau du produit de ces gènes et de la régulation de leur expression. Plus on en connaît sur une maladie, mieux on peut la combattre ou du moins limiter ses conséquences.

L'étude du transcriptome et des régulations transcriptionnelles sont donc un des aspects très importants de l'étude des pathologies. De plus en plus de publications portent par exemple sur l'étude des Acides RiboNucléiques (ARN, ou RNA en anglais) non-codants comme les microRNA, ou bien encore les « *Long intergenic non-coding RNA* » (lincRNA), et leurs implications dans des pathologies (voir partie 1.5).

1.2 Le transcriptome

Le transcriptome est l'ensemble des ARN issus de l'expression d'une partie des gènes du génome d'un type cellulaire ou d'un tissu à un moment et dans des conditions données. Initialement focalisée sur l'ensemble des ARN messagers, cette définition a été étendue aux ARN non codants comme les microRNA, les ARN ribosomaux, les ARN de transfert ou bien encore les lincRNA suite à la mise en évidence récente, grâce au séquençage à haut débit, que plus de 80% des nucléotides d'un génome peuvent être transcrit. Cependant, cette transcription généralisée n'implique pas nécessairement une fonction associée à chacun des produits. Nous retiendrons donc le terme de transcriptome pour désigner l'ensemble des transcrits, codant ou non codant, associés à une fonction dans un organisme.

La caractérisation et la quantification du transcriptome dans un modèle biologique (tissu, organisme, cellule ou lignée cellulaire) permettent d'identifier les gènes transcrits dans un contexte donné et ainsi de déterminer les mécanismes de régulation d'expression des gènes (co-expression) et de définir leurs réseaux de régulation (voies de signalisation impliquées). La meilleure connaissance du niveau d'expression d'un gène dans différentes situations constitue une avancée vers la compréhension de sa fonction, mais également vers le criblage de nouvelles molécules et l'identification de nouveaux médicaments et de nouveaux outils de diagnostic.

Introduite dans les années 1980, la technique à haut débit des **puces à ADN** permet de mesurer simultanément le niveau d'expression d'un large ensemble d'ARN messagers contenus dans un échantillon, ce qui en fait un outil de choix pour l'étude du transcriptome. Cette méthode est de nos jours encore couramment utilisée dans les laboratoires de recherche pour diverses applications telles que l'identification de cibles thérapeutiques, de biomarqueurs ou de voies de signalisation impliquées dans une pathologie, la caractérisation des mécanismes de résistance à une drogue, ou encore l'identification de signatures transcriptionnelles dans divers

contextes biologiques. Son utilisation intensive a entraîné le développement de nombreuses technologies permettant l'acquisition des données. De plus, elle a nécessité le développement de nombreux outils et méthodes bioinformatiques et statistiques dédiés au traitement de la masse de données ainsi obtenue.

Actuellement, avec le développement du séquençage à très haut débit, de nouvelles techniques d'études du transcriptome ont vu le jour : le RNA-seq et le SAGE-seq. Ces techniques détaillées plus loin ne sont pas celles que j'ai utilisées durant ma thèse.

1.2.1 Principe des puces à ADN

Le principe des puces à ADN, encore appelées biopuces ou *microarrays*, repose sur l'hybridation d'un échantillon de séquences d'ADN ou d'ARN complémentaires (ADNc, ARNc) marqué avec des brins d'ADN plus courts, complémentaires, ou des oligonucléotides synthétiques, fixés sur un support solide. Les premières puces, créées au milieu des années 1980, utilisaient une membrane de nylon et un marquage radioactif (appelées parfois *macroarrays* par opposition aux *microarrays* actuels). Elles ont ensuite été supplantées dans les années 90 par la technologie sur lame de verre avec un marquage fluorescent. La miniaturisation sur support solide, l'utilisation de marqueurs fluorescents et les progrès de la robotique permettent aujourd'hui de fabriquer des puces comportant une très haute densité d'unités d'hybridations ou « *spots* ». Chaque spot est constitué de sondes c'est-à-dire d'oligonucléotides (d'une longueur de quelques dizaines de nucléotides) ou de produits PCR (pour « *Polymerase Chain Reaction* », ADNc d'une longueur de quelques centaines de nucléotides). Ces sondes correspondent à des séquences d'ADN spécifiques d'un transcrit codants connu ou prédit. Les oligonucléotides synthétisés sont issus de banques de données telles que GenBank ou dbEST et correspondent donc à des séquences non redondantes spécifiques d'un transcrit donné.

A l'heure actuelle, les avancées technologiques ayant permis une augmentation spectaculaire de la densité des spots, les puces les plus fréquemment utilisées sont pangénomiques, c'est-à-dire que les sondes interrogent tous les transcrits connus d'un génome ainsi que quelques séquences non annotées. En plus des séquences d'ARN messagers, elles comportent parfois celles correspondant à des lincRNAs.

On distingue différents formats et types de puces (encore appelées plateformes) selon la densité des spots, la nature et le mode de fabrication des sondes (synthèse *in situ* par photolithographie ou impression « jet d'encre »), la nature des cibles, les méthodes d'hybridation et le champ d'application. Plusieurs sociétés commerciales ont ainsi développé des puces à ADN parmi lesquelles Agilent Technologies, Affymetrix, GE Healthcare, Life Technologies (Applied Biosystems) ou encore Illumina. Le choix de l'Unité UMR_S 928 TAGC s'étant porté en 2007 sur l'installation d'une plateforme transcriptome commerciale Agilent^{TM}, celle-ci sera décrite ci-après.

1.2.2 Cas particulier des puces à ADN de technologie Agilent™

La technologie développée dans les années 90 par Agilent Technologies utilise un support d'hybridation rigide de type lame de verre, qui permet le dépôt d'une densité élevée de sondes grâce à une technique d'impression de type « jet d'encre ». Dans un premier temps, des ARNc marqués avec un fluorochrome obtenues à la suite d'une transcription inverse et d'une transcription *in vitro* des ARN de l'échantillon. Ces ARNc sont ensuite incubés avec la puce afin de permettre leur hybridation avec les sondes présentes sur la celle-ci (Figure 1.1).

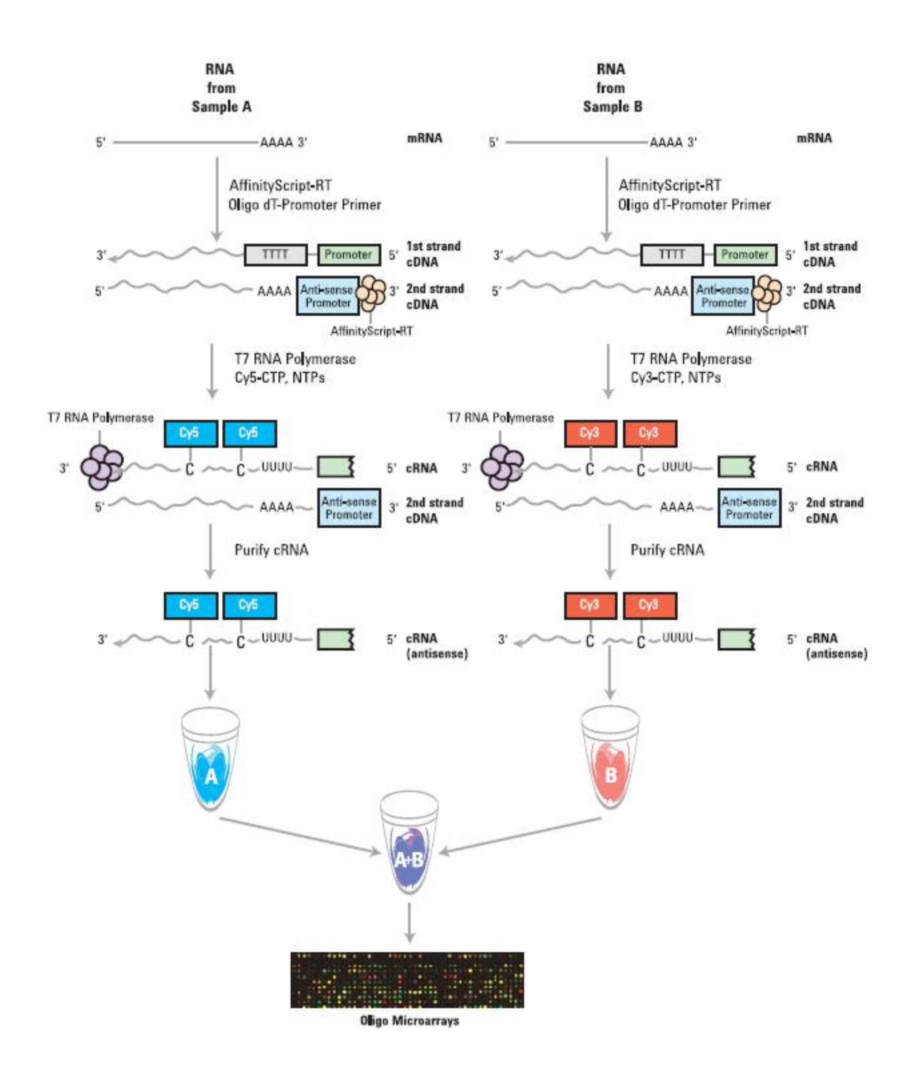


FIGURE 1.1 – Procédure d'amplification des ARNc pour une expérience bi-canale : pour une expérience mono-canale, seuls les échantillons marqués au Cy3 (« B ») seront utilisés (Extrait du manuel d'Agilent™« One-Color Microarray-Based Gene Expression Analysis (Low Input Quick Amp Labeling) Protocol »)

Ces puces comportent des sondes à longs oligonucléotides (60 nucléotides) contrairement à la technologie Affymetrix™dans laquelle les sondes sont beaucoup plus courtes (25 nucléotides). L'acquisition des données de fluorescence pour chaque spot est réalisée avec un scanner Agilent (G2565CA) équipé d'un système de balayage laser permettant d'exciter chaque fluorochrome à sa longueur d'onde spécifique. L'image de la puce générée est par la suite traitée par un logiciel de quantification afin de calculer l'intensité de chaque spot, c'est à

dire le niveau d'expression de chaque transcrit représenté sur la puce.

De nombreux formats de puces pangénomiques sont disponibles en version mono et bi-canale (j'emploierai par la suite les termes anglais **one-color** et **two-colors** qui sont préférentiellement utilisés dans le milieu scientifique), pour les principaux organismes modèles tels que l'homme, la souris, le rat et la levure. Leurs identifiants sont composés du nombre d'échantillons multiplié par le nombre de sondes en milliers (« k ») voire aujourd'hui en millions (« m »). Ces formats sont différents en fonction du type d'impression : les SurePrint HD (8x15k, 4x44k, 2x105k, 1x244k) mais également la nouvelle génér ation de puces contenant des lincRNAs, les SurePrint G3 (8x60k, 4x180k, 2x400k, 1x1m). Il est également possible d'obtenir des puces à ADN à façon pour étudier le transcriptome d'espèces atypiques à l'aide du logiciel eArray.

Il existe également d'autres types de puces, chacune ayant une application bien définie, telles que : les puces CGH (« *Comparative Genomic Hybridation* ») pour étudier les amplifications et les délétions génomiques sur l'ensemble du génome ([Solinas-Toldo *et al.*, 1997], [Snijders *et al.*, 2001]), les puces recouvrant l'entièreté du génome par fragments chevauchants (« *tiling arrays* ») pour des applications comme le ChIP-on-chip (voir partie 5.1.1), les puces de génotypage SNP (« *Small Nucleotide Polymorphism* ») permettant l'analyse des polymorphismes ([Pastinen *et al.*, 2000]).

Une expérience de puces à ADN se déroule selon les étapes suivantes : conception du plan d'expérience, marquage et hybridation, acquisition et traitement des données, analyse et interprétation des résultats (Figure 1.2). **Les particularités de l'acquisition, de la correction, de la normalisation des données de technologie AgilentTM seront présentées et discutées dans les chapitre 2 et 3 de ce manuscrit. La méta-analyse de données de puces sera présentée au travers du projet TranscriptomeBrowser développé au TAGC et auquel j'ai contribué (voir Chapitre 4).**

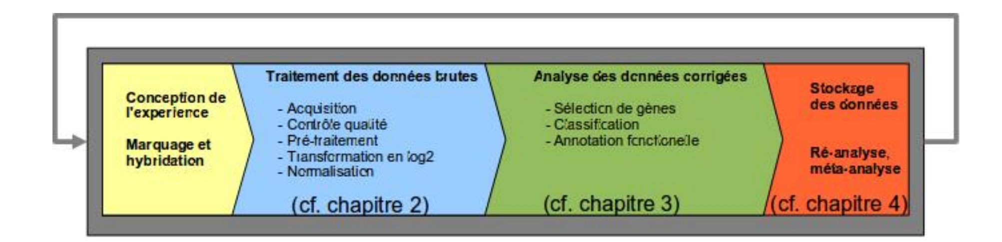


FIGURE 1.2 – Déroulement d'une expérience de puces à ADN, du dessin de l'expérience au stockage des données, en passant par leur traitement et leur analyse. Chacune de ces étapes a nécessité le développement d'outils bioinformatiques qui seront détaillés dans les chapitres 2 à 4 de ce manuscrit

1.3 Régulation de l'expression des gènes

La **régulation de l'expression des gènes** est un processus important chez les organismes vivants. Dans le cas des organismes pluricellulaires, toutes les cellules possèdent a priori le même patrimoine génétique. Le nombre de gènes du génome humain est fixe, c'est la régulation qui permet aux cellules d'exprimer différemment leurs gènes et de se différencier au cours de l'embryogenèse pour donner différents tissus. De même, ces tissus possèdent des caractéristiques et une régulation particulière au niveau des gènes et de la structure même de la conformation de l'ADN.

La régulation de l'expression des gènes comprend l'ensemble des mécanismes de régulation mis en oeuvre pour passer de l'information génétique incluse dans une séquence d'ADN à un produit fonctionnel (ARN ou protéine). Elle comporte plusieurs niveaux : **transcriptionnel, post-transcriptionnel, traductionnel et post-traductionnel**.

Seules les régulations transcriptionnelles seront décrites ci-après, compte tenu de mes travaux de thèse portant sur l'étude de l'expression des gènes et de leur régulation par l'analyse de la fixation de facteurs de transcription.

L'expression des gènes est le résultat de l'interaction de plusieurs processus : (1) la **transcription basale** par les ARN polymérases et les facteurs de transcription généraux, (2) la modulation de celle-ci par des f**acteurs de transcription séquences-spécifiques**, (3) la **dynamique de la chromatine** par la modulation des marques épigénétiques, et (4) la régulation par les **ARN non codants**.

1.3.1 La transcription basale

La transcription basale de l'ADN en ARN s'effectue sous l'influence des ARN polymérases et de nombreux facteurs de transcription généraux. Les ARN polymérases sont dits ADN dépendants et leur type permet la transcription d'ARN différents. Ainsi, le type I est à l'origine des ARN ribosomiques, alors que le type II est à l'origine des ARN messagers, de la plupart des small nuclar RNA (snRNA), des small nucleolar RNA (snoRNA) et des microRNA ([Kornberg, 1999], [Sims *et al.*, 2004]) ; enfin, le type III est à l'origine de la synthèse des ARN de transfert ainsi que de l'ARN ribosomique 5S. Quant aux facteurs de transcription généraux, comme la famille TFII (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH et TFIIS) ([Lee & Young, 2000]), ils sont requis pour permettre le recrutement de l'ARN polymérase II (PolII) aux promoteurs, formant ainsi le complexe de pré-initiation de la transcription ([Orphanides *et al.*, 1996]) (Figure 1.3).

Les gènes des eucaryotes possèdent des séquences régulatrices présentes à proximité du site d'initiation de la transcription (ou TSS pour « *Transcription Start Site* ») constituant le promoteur proximal. Celui-ci est le lieu de formation du complexe de pré-initiation de la transcription. La modulation de la transcription de l'ADN par la PolII est effectuée par des facteurs de transcription spécifiques. Leur liaison non-covalente au niveau de l'ADN s'effectue

au niveau de sites spécifiques, appelés sites de fixation de facteurs de transcription (TFBS, « *Transcription Factor Binding Site* »), afin d'activer ou d'inhiber l'expression d'un gène donné.

1.3.2 Les séquences régulatrices et les facteurs de transcription séquences-spécifiques

La modulation de l'expression des gènes est rendue possible par l'assemblage de protéines, tel que les **facteurs de transcription** liés à l'ADN au niveau des séquences régulatrices. Ce contrôle spatiotemporel de l'expression des gènes au sein de l'organisme permet la mise en place et le maintien de la spécificité tissulaire, impliquant de nombreuses voies de signalisation et réseaux de régulation transcriptionnelle ([Naef & Huelsken, 2005], [Zhang *et al.*, 2004], [Visel *et al.*, 2009*a*]).

Ces régions régulatrices sont de plusieurs types : (1) les **promoteurs** lorsqu'ils sont situés à proximité du site d'initiation de la transcription de la région codante, (2) les **enhancers** quand ils se localisent à distance du site d'initiation de la transcription et qu'ils potentialisent l'action du promoteur, (3) des **silencers** lorsqu'ils se situent comme les enhancers à distance du gène mais qu'ils le répriment, et (4) les **insulateurs** qui correspondent à une séquence régulatrice affectant l'interaction entre l'enhancer et le promoteur ([Geyer & Corces, 1992], [Kellum & Schedl, 1992]). Ces régions régulatrices sont particulièrement conservées lors de l'évolution, car elles sont composées de courtes séquences de 6 à 15 paires de bases (pb), appelées **éléments régulateurs** (RE, « *Regulatory Element* ») permettant le recrutement spécifique des facteurs de transcription sur l'ADN (Figure 1.3).

Les facteurs de transcription n'agissent pas de manière indépendante, mais forment des complexes avec d'autres facteurs de transcription et des cofacteurs protéiques, comme c'est également le cas pour les facteurs de transcription généraux ([Fedorova & Zink, 2008], [Ravasi *et al.*, 2010]). Ces facteurs de transcription se lient à leur sites de liaison spécifiques souvent regroupés en modules cis-régulateurs.

1.3.3 La chromatine : histones et marques épigénétiques

La **chromatine** est composée de l'enroulement de la double-hélice d'ADN chromosomique autour des nucléosomes d'histones et des protéines non histones. La chromatine est ainsi un **polymère de nucléosomes** dont le degré de condensation affecte l'accessibilité de la machinerie transcriptionnelle à l'ADN. Les nucléosomes sont des **octamères d'histones** constitués de deux hétérodimères H2A/H2B et H3/H4 autour desquels 146 paires de bases (pb) d'ADN sont enroulés (Figure 1.4 A). L'histone linker H1, est localisée entre 2 nucléosomes et permet une condensation supplémentaire de la chromatine. La chromatine est présente sous deux états : l'hétérochromatine sous forme compactée et inaccessible aux activités enzymatiques, et l'euchromatine sous forme décondensée laissant ainsi accessible l'ADN à la machinerie des ARN polymérases. L'accessibilité de la chromatine est un excellent indicateur de la capacité de liaison des facteurs de transcription et de la dynamique des nucléosomes participant à la

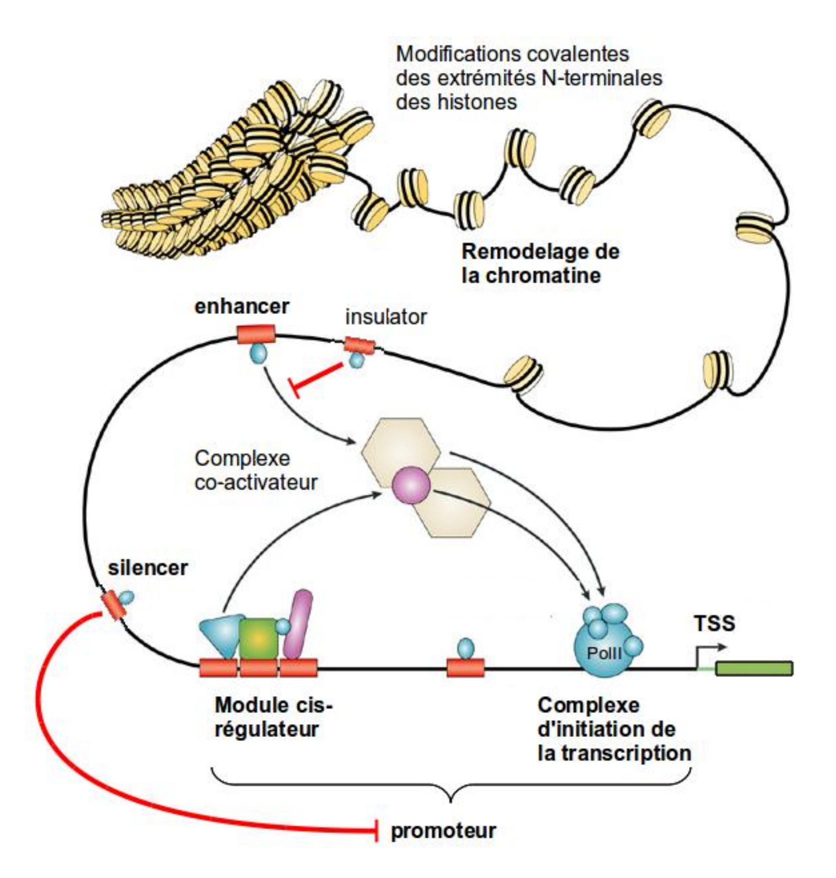


FIGURE 1.3 – Représentation schématique des régions régulatrices permettant la modulation transcriptionnelle de l'expression des gènes (Adapté de [Wasserman & Sandelin, 2004])

régulation de l'expression des gènes.

La modulation de cette accessibilité est directement dépendante de la structure dynamique de, qui par intégration de signaux spécifiques joue un rôle important dans la régulation de l'expression des gènes ([Li *et al.*, 2007]). Le remodelage de la chromatine permettant les principaux mécanismes cellulaires exige une grande plasticité controlée par des modifications biochimiques de sa structure à l'aide des mécanismes épigénétiques tels que les modifications des histones, leur substitution par des variants et la méthylation de l'ADN au niveau des dinucléotides CG souvent concentrés dans des courtes régions (>200 bp, [CpG]>60%), appelées ilots CpG (Figures 1.4 et 1.5). En effet chez l'homme, on observe une sous-représentation globale des dinucléotides CpG (environ 20% de la fréquence attendu) et une surreprésentation locale à proximité des régions promotrices et des enhancers (29 000 ilots CpG prédits dans l'ensemble du génome humain).

La **méthylation de l'ADN** est une modification épigénétique transmissibles. Sa présence est généralement associée à la répression de la transcription gèniques. Elle se localise principalement au niveau des ilots CpG à proximité des gènes (Figure 1.5).

Le degré de condensation de la chromatine est contrôlé par des **modifications des extrémités N-terminales des histones**, comme des phosphorylations, acétylations, méthylations,

ubiquitinations, sumoylations ([Kouzarides, 2007]) (Figure 1.4 B et C). Toutes ces modifications sont catalysées par des enzymes spécifiques. Les modifications covalentes des histones agiraient soit directement en modifiant la compaction de l'enroulement d'ADN autour des nucléosomes, soit indirectement en constituant des « marques » permettant le recrutement de protéines capables de remodeler localement la structure de la chromatine. Le modèle des modifications covalentes des histones agissant comme un code (le « code des histones ») a été proposé par Strahl et Allis en 2000 ([Strahl & Allis, 2000], [Jenuwein & Allis, 2001]). Ce code est loin d'être universel. Il serait plus ou moins spécifique selon les gènes et les cellules considérés, mais semblerait être évolutivement stable au sein des mammifères ([Lee & Mahadevan, 2009]). Cependant, le rôle des marques épigénétiques dans le maintien de l'identité cellulaire n'est pas encore clairement défini ([Natoli, 2011]).

L'effet fonctionnel des principales marques d'histones dépend au moins en partie de leur localisation. Ainsi, l'étude de leur profil le long des gènes mais également de leurs séquences régulatrices a pu montrer que les promoteurs actifs présentent des modifications du type H3K4me3 et H3K27ac, tandis que les enhancers actifs seraient plutôt sujets à H3K4me1 et H3K27ac. Les gènes transcrits possèderaient des modifications de type H3K36me3, alors que l'hétérochromatine inaccessible aux éléments de régulation présenterait des marques H3K9me3 et H3K27me3 ([Visel *et al.*, 2009*b*], [Heintzman *et al.*, 2009]).

Il existe également des **variants d'histones** qui jouent des rôles majeurs dans différents processus tels que la réparation de l'ADN ([Klose & Zhang, 2007], [Billon & Côté, 2011]), l'organisation centromérique ([Foltz *et al.*, 2009]), l'inactivation du chromosome sexuel ([Fernandez-Capetillo *et al.*, 2003]) et une condensation spécifique des cellules gamètes mâles ([Okada *et al.*, 2005], [Govin *et al.*, 2004]). Ces variants ont une séquence qui diffère de celle des histones conventionnelles sur quelques résidus seulement ou sur des portions plus importantes de la protéine.

1.3.4 Les ARN non codants

De récentes analyses transcriptomiques à très haut débit ont mis en évidence que plus de 90% du génome est transcrit, mais que seuls 1 à 2 % de ces transcrits coderaient pour des protéines ; les autres constitueraient une catégorie de transcrits appelés ARN non-codants (ncRNA pour *non-coding RNA*). Parfois bien conservés lors de l'évolution, ce qui suggère une importance fonctionnelle, ils sont cependant généralement moins fortement exprimés que les ARN messagers.

Ces ncRNA peuvent être divisés en 2 groupes : les ncRNA d'infrastructure, incluant les ARN ribosomiques, les ARN de transfert et les petits ARN de type small nuclear, et les ncRNA de régulation, comme les micro ARN (miRNA), les *small interfering RNA* (siRNA) et les *long non-coding RNA* (lncRNA) ([Ponting *et al.*, 2009]). Au-delà de leur rôle dans la dégradation d'un ARNm cible, les miRNA et les siRNA et les ont ainsi été identifiés, ainsi que le lncRNA, comme pouvant jouer un rôle dans la régulation de l'expression des gènes par le ciblage des promoteurs et l'activation de la traduction ([Krol *et al.*, 2010], [Place *et al.*, 2008], [Kim *et al.*, 2008]).

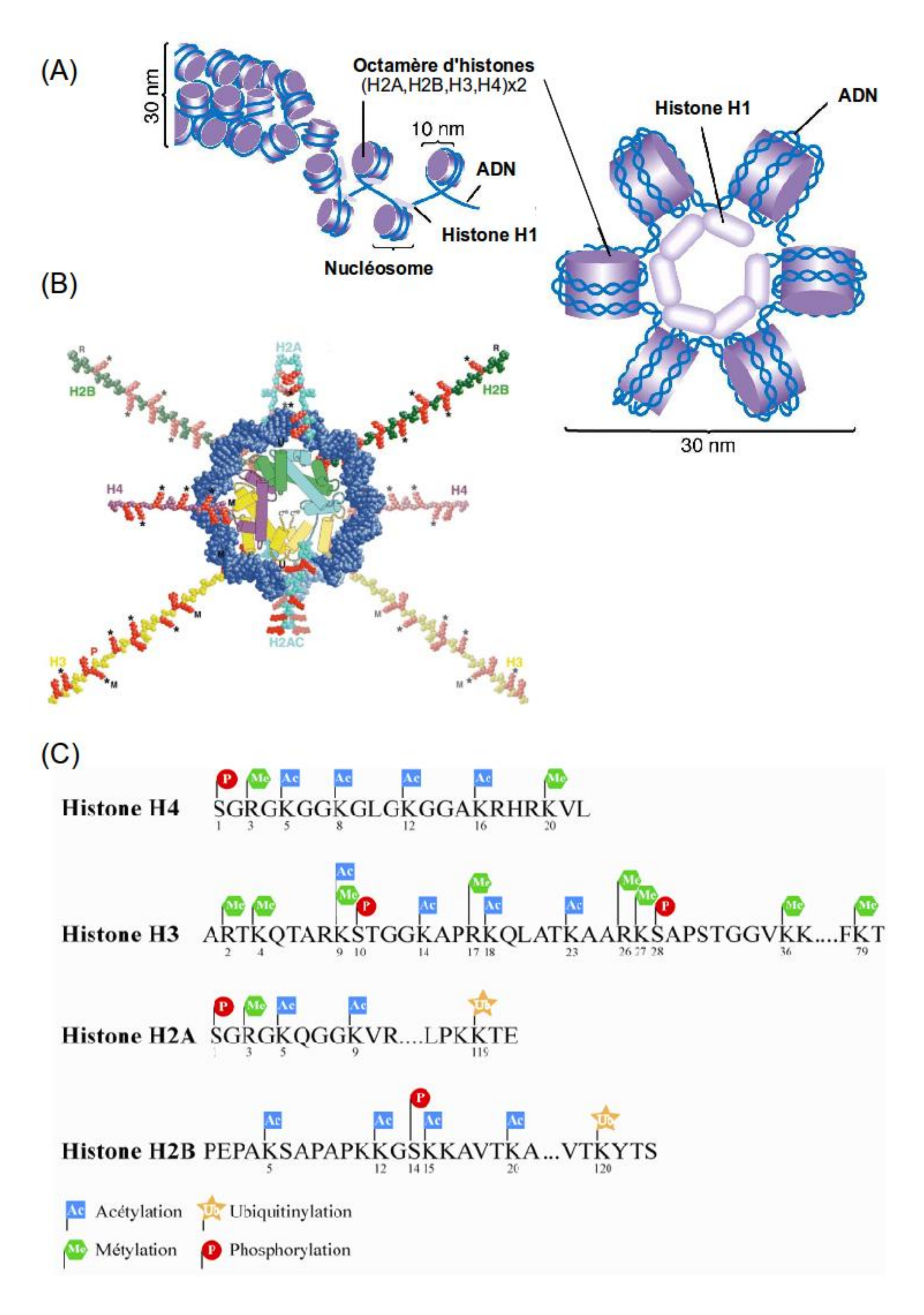


Figure 1.4 – Représentation des modifications covalentes d'histones avec en (A) la structure de la chromatine avec ces octamères d'histones (adapté de http://www.mun.ca/biology/scarr/Histone_Protein_Structure.html), en (B) la structure tridimensionnelle d'un nucléosome avec le positionnement des principales modifications d'histones (extrait de [Wolffe & Hayes, 1999]) et enfin en (C) les diverses modifications N-terminales des histones H2A, H2B, H3 et H4 (adapté de [Lacoste & Côté, 2003])

La majorité des lncRNA sont de taille supérieure à 200nt. Récemment, trois nouvelles classes de lncRNA ont été décrites : les *long intergenic non-coding RNA* (lincRNA), *enhancer RNA* (eRNA) et les *promoter-associated RNA* (PAR) ([Kim *et al.*, 2010], [De Santa *et al.*, 2010], [Ø rom *et al.*, 2010]).

Les lincRNA auraient une signature chromatinienne identique à celle des gènes actifs. Ils possèdent donc des marques épigénétiques telles que H3K4me3 au niveau de leur promoteur et H3K36me3 le long de la région transcrite. On pense qu'ils jouent le rôle de guide des modifications de la chromatine participant ainsi à l'établissement d'un état épigénétique spécifique pour chaque type cellulaire ([Khalil *et al.*, 2009], [Guttman *et al.*, 2009]).

La taille des eRNA varie entre 100 et 900 nt ([De Santa *et al.*, 2010], [Ø rom *et al.*, 2010]). Contrairement aux lincRNA, les eRNA possèdent les caractéristiques épigénétiques spécifiques des enhancers, d'où leur nom. En effet, ceux-ci sont enrichis en marques de H3K4me1 et sont transcrits par l'ARN polymérase II (contrairement aux autres ARN non codants transcrits par l'ARN polymérase III), et de co-régulateurs tels que le co-activateur p300. En revanche, ils sont faiblement enrichis en H3K4me3.

Enfin, les PAR sont une catégorie de plus petite taille comprise entre 16 et 200nt. Ces ncRNA sont caractérisés par leur localisation ; certains sont exprimés à proximité des TSS, alors que d'autres le sont au niveau des promoteurs. Un nombre croissant d'études semble indiquer que les PAR jouent un rôle dans la régulation de l'expression des gènes aussi bien dans leur activation que dans leur répression ([Morris *et al.*, 2008], [Wang *et al.*, 2008], [Kaikkonen *et al.*, 2011]).

1.3.5 Epigénétique et épigénomes

L'information contenue dans le génome est ainsi spécifiquement régulée par les marques épigénétiques de manière spatio-temporelle pour permettre aux processus biologiques de se produire dans un type cellulaire donné et au stade développemental approprié ; tel est le rôle de la **régulation épigénétique**. Ce terme, dont la définition initiale fut introduite en 1942 par Conrad H. Waddington, définit les modifications transmissibles et réversibles de la chromatine ne s'accompagnant pas de changements de la séquence nucléotidique de l'ADN.

Il est cependant important de ne pas confondre épigénétique et épigénome. En effet, l'épigénome est l'état épigénétique de la cellule. Un épigénome fait donc référence aux caractéristiques épigénétiques d'une cellule donnée telles que la méthylation de l'ADN, les modifications d'histones et l'accessibilité de la chromatine permettant l'accès au génome et ainsi l'expression d'ARN messagers et non-codants ([Bernstein *et al.*, 2010]). Chaque type cellulaire à un état de différenciation donné possède ainsi son épigénome, qui définit son programme d'expression génique.

Une meilleure compréhension des mécanismes de régulation épigénétiques et des épigénomes a été rendue possible par le développement du séquençage à très haut débit, décrit dans la partie suivante de ce chapitre. L'utilisation de ces approches dans une variété des tissus a ainsi

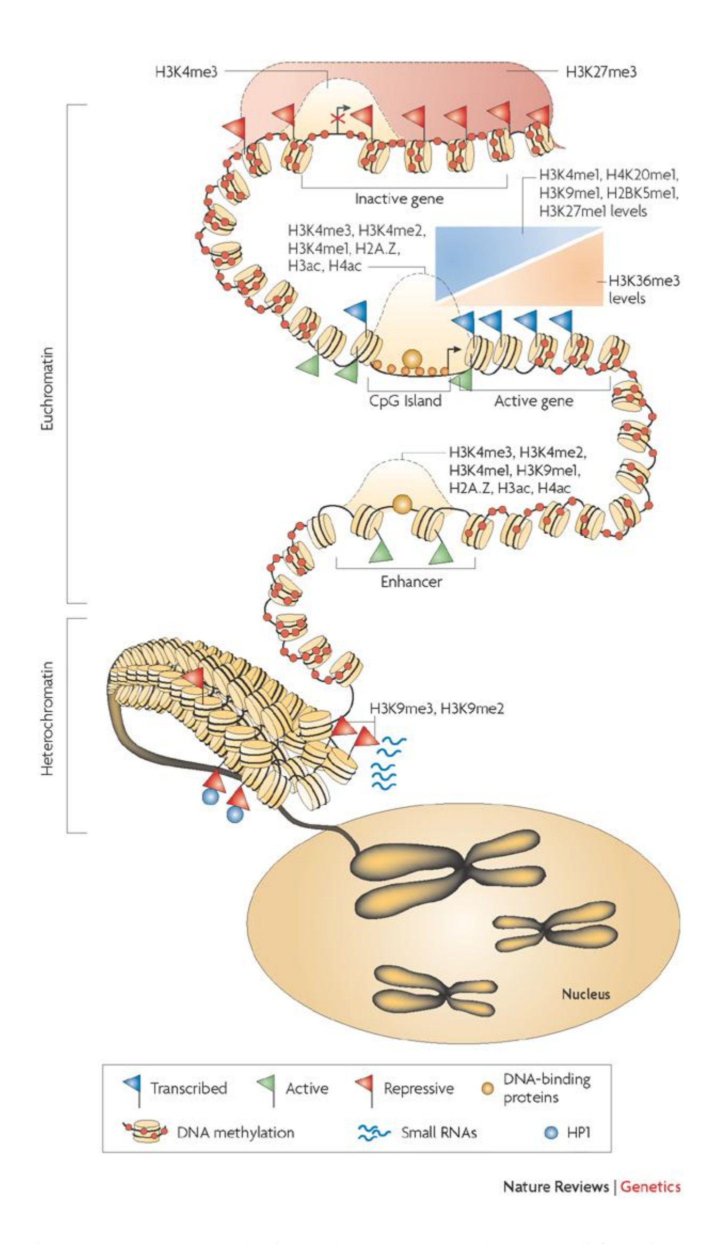


Figure 1.5 – Interaction de la méthylation de l'ADN, des modifications d'histones, du positionnement des nucléosomes et des autres facteurs permettant la régulation de l'expression des gènes comme des facteurs de transcription et les *small RNA*. Les régions d'euchromatine sont marquées par H3K9me2 et H3K9me3, qui servent à la liaison de HP1 (Heterochromatic protein 1). Les *small RNA* sont impliqués dans le maintien de l'hétérochromatine.La méthylation de l'ADN est présente tout le long du génome mais est absente généralement au niveau des régions régulatrices des gènes actifs ou activables). La modification H3K27me3 marque les gènes inactifs tandis que H3K4me3, H3K4me2, H3K4me1, l'acétylation des histones et le variant de l'histone H2A (H2A.Z) marquent la région d'initiation de la transcription des gènes actifs. Les mono-méthylations de H3K4, H3K9, H3K27, H4K20 et H2BK5 sont localisées au niveau des régions transcrites avec un pic en 5' du gène, alors que H3K36me3 marque également les régions transcrites mais avec un pic en 3' du gène (Extrait de [Schones & Zhao, 2008])

permis d'avoir une vue globale à l'échelle du génome de l'activité des éléments cis-régulateurs, de la fonction des facteurs de transcription et des processus épigénétiques impliqués dans le contrôle de l'expression des gènes.

1.4 Les techniques de séquençage à très haut débit

Les techniques de séquençage à très haut débit (HTS pour « *High Thoughput Sequencing* »), communément, et improprement, appelées NGS pour « *Next Generation Sequencing* », connaissent un développement spectaculaire depuis leur apparition commerciale au début 2006 ([Margulies *et al.*, 2005], [Shendure *et al.*, 2005], [Hutchison, 2007], [Chan, 2005]). Elles constituent la troisième génération de séquençage après les méthodes de **Sanger**, et de **Maxam et Gilbert** en 1977 et le **pyroséquencage** en 1988.

Le séquençage de l'ADN a été inventé dans la deuxième moitié des années 1970. Deux méthodes ont été développées indépendamment : une méthode de dégradation chimique sélective par l'équipe de Walter Gilbert ([Maxam & Gilbert, 1977]) et une méthode de synthèse enzymatique sélective par Frederick Sanger ([Sanger *et al.*, 1977], [Prober *et al.*, 1987]). Puis le pyroséquençage a été développé et reste encore aujourd'hui une technique très utilisée car elle est plus rapide que les méthodes classiques ([Hyman, 1988], [Ronaghi *et al.*, 1998]). C'est l'une des raisons pour lesquelles cette technique a été choisie par une des technologies du HTS qui sera brièvement décrite plus tard dans ce manuscrit.

À la différence des premières générations de séquenceurs à capillaires, ces les séquenceurs à très haut débit actuels permettent le **séquençage massif en parallèle de plusieurs millions de fragments d'ADN**, ceci très rapidement et en diminuant les coûts et avec une moindre quantité de matériel biologique. Cette avancée a nécessité des développements technologiques constant tant au niveau biologique (automatisation, révision et amélioration des réactifs, des protocoles) qu'au niveau informatique (algorithmes, logiciels, ferme de calcul, mémoire, stockage).

Grâce aux techniques du HTS, des études qui n'étaient pas envisageables pour diverses raisons (trop longues, pas assez de matériel biologique, trop coûteuses) ont pu être réalisées ([Hillier *et al.*, 2008], [Srivatsan *et al.*, 2008]). Il est dorénavant possible de séquencer plusieurs centaines de giga-bases (Gb) du génome avec une couverture suffisante permettant des études de liaison génétiques telles que la recherche des polymorphismes spécifiques par exemple, des SNP, pour *Single Nucleotide Polymorphisms*), présents chez plusieurs patients. Jusque-là en effet, l'expérimentateur était contraint de sélectionner des gènes d'intérêt et de séquencer spécifiquement de courtes régions de quelques centaines de bases préalablement identifiées. Ainsi, le Projet Génome Humain, entrepris en 1990 avec pour mission de déterminer la séquence du génome humain par la méthode de Sanger, n'a pu être achevé qu'en avril 2003, avec tout de même deux ans d'avance sur la date prévue. La réalisation de ce projet a nécessité l'intervention d'un nombre important de biologistes pour le séquençage et de bioinformaticiens pour l'assemblage des séquences ainsi que pour le développement et l'utilisation de puissants moyens informatiques.

Avec les HTS, le séquençage *de-novo* ou le re-séquençage du génome humain est possible en seulement quelques jours dans les grands centres de séquençage (Figure 1.6 C). En effet, l'automatisation de tous les processus expérimentaux et l'utilisation d'un pipeline d'analyse automatique permettent un séquençage et une analyse extrêmement rapides des échantillons. De plus, des compagnies dédiées à un type d'analyse spécifique ont été créées ; elles fonctionnent 24h/24 et 7j/7 et produisent de l'ordre d'un milliard de séquences par jour (par exemple, Complete Genomics ou le BGI).

1.4.1 Principes du séquençage à très haut débit

Depuis 2005, diverses technologies ont été mises au point pour permettre le séquençage à très haut débit de plusieurs millions de séquences d'ADN en parallèle. On peut distinguer trois technologies principales proposées par différents fournisseurs, chacune possédant des caractéristiques particulières et ayant recours à des techniques spécifiques : Roche Diagnostics /454 Life Sciences, Illumina/Solexa et Life Technologies/Applied Biosystems (ABI) (Table 1.1) ([Metzker, 2010], [Suzuki *et al.*, 2011]).

A l'heure actuelle, plus de 1 800 séquenceurs de nouvelle génération ont été vendus à travers le monde (Figure 1.6 A), 93,3% appartiennent à l'une de ces quatre technologies, dont plus de la moitié correspondant à un modèle Illumina (Figure 1.6 B). Des centres de séquençage de renommée internationale se sont équipés d'un nombre important de séquenceurs à très haut débit (Figure 1.6 C) ([Hum, 2010]).

La gamme de séquenceurs en développement s'étend très rapidement ; ainsi, seules les techniques et les modèles les plus couramment utilisés seront décrits.

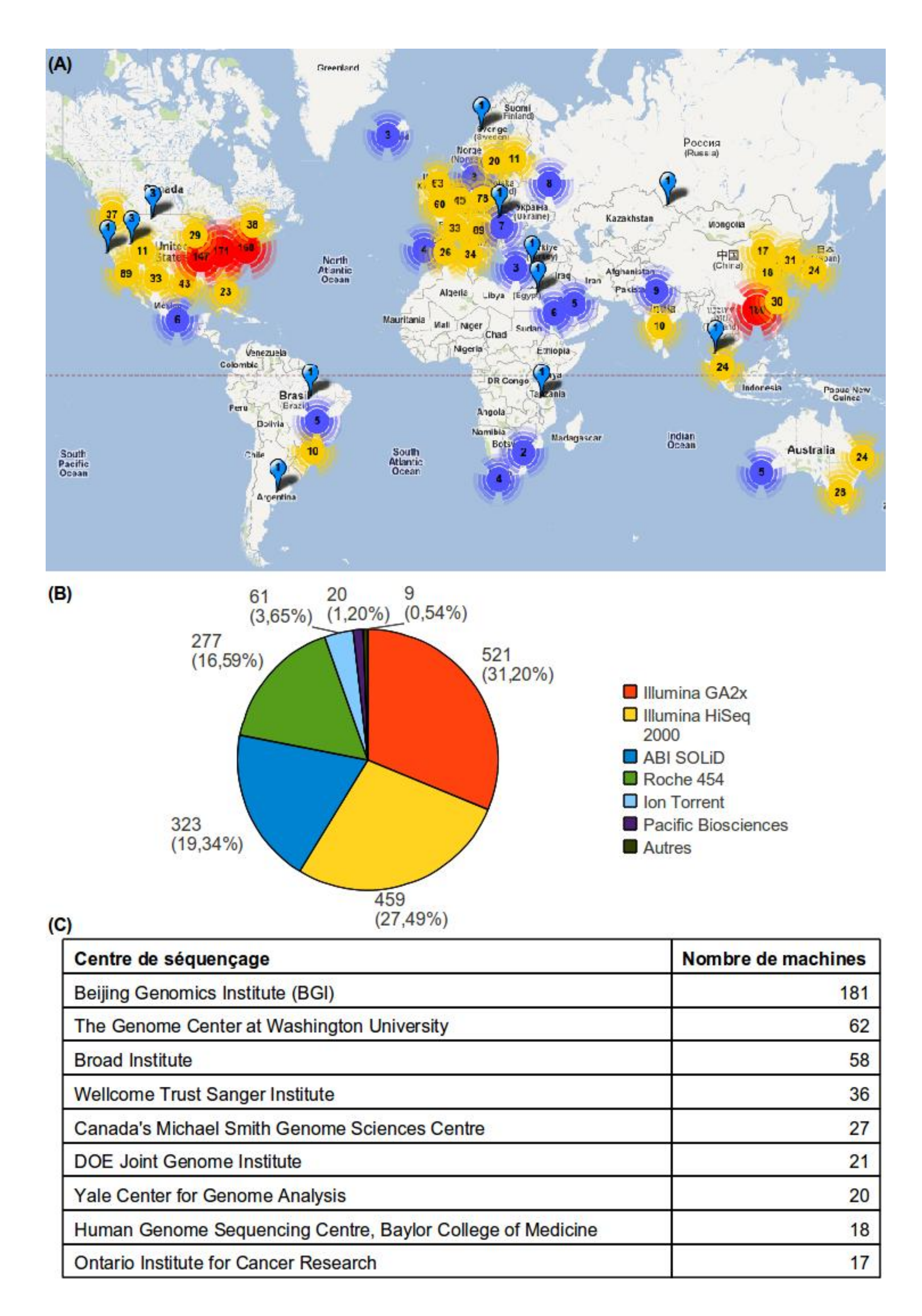
Pendant ma thèse, la plateforme TGML a fait le choix de s'équiper d'un séquenceur à très haut débit. J'ai ainsi participé aux discussions avec les divers fournisseurs. C'est pourquoi je présenterai ci-après les trois principales chimies de séquençage. La plateforme s'est finalement équipé en avril 2009 d'un séquenceur à très haut débit de type SOLiD™. J'ai participé à de nombreuses collaborations sur l'analyse de données provenant d'expériences de Chromatine ImmunoPrécipitation (ChIP-seq). Leurs analyses au niveau de la plateforme a nécessité le développement d'un pipeline de traitement des données spécifique du séquenceur SOLiD™. Tous ces résultats seront présentés dans le chapitre 5 de ce manuscrit.

1.4.1.1 Chimie

Bien que les trois principales technologies de séquençage possèdent chacune une chimie et des caractéristiques de séquençage différentes (Table 1.2), elles permettent toutes de séquencer, en parallèle, des fragments d'ADN obtenus après une étape d'amplification (Figure 1.7) ([Suzuki *et al.*, 2011], [Borgström *et al.*, 2011], [Shendure *et al.*, 2005]).

Société	Support	Amplification	Technique de séquençage	Modèle	Année
Life Technologies	Lame de verre	Emulsion PCR	Ligation	SOLiD v2	2007
				SOLiD v3	2008
				SOLiD v3.5	2009
				SOLiD v4	2010
				SOLiD 5500 XL	2011
	Puce semi-conductrice	-	Différentiel de potentiel électrique	Ion Torrent *	2010
Illumina	Lame de verre	Pontage sur phase solide	Synthèse	GA I	2007
				GA IIx	2008
				HiScanSQ	2009
				HiSeq 1000	2009
				HiSeq 2000	2010
				MiSeq *	2011
Roche Diagnostics	Plaque à pico-titration (« PicoTiterPlate », PTP)	Emulsion PCR	Pyroséquencage	GS20	2006
				GS FLX	2007
				GS FLX Titanium	2008
				GS Junior *	2011

Table 1.1 – Tableau comparatif des principales technologies de séquençage à très haut débit Les cellules grisées correspondent à des technologies à très haut débit de dernière génération (NGS) qui seront décrites plus tard dans ce manuscrit. Les « * » pointent les modèles de séquenceurs « de paillasse » (de petite taille) à faible débit, mais très rapides



Centre de séquençage	Nombre de machines
Beijing Genomics Institute (BGI)	181
The Genome Center at Washington University	62
Broad Institute	58
Wellcome Trust Sanger Institute	36
Canada's Michael Smith Genome Sciences Centre	27
DOE Joint Genome Institute	21
Yale Center for Genome Analysis	20
Human Genome Sequencing Centre, Baylor College of Medicine	18
Ontario Institute for Cancer Research	17

FIGURE 1.6 – Répartition des différentes technologies de séquençage à très haut débit dans le monde en décembre 2011. (A) Répartition géographique, (B) Distribution en nombre et en pourcentage des principaux modèles de séquenceurs à très haut débit (nombre total : 1670), et (C) Principaux centres de séquençage (source : http://pathogenomics.bham.ac.uk/hts)

Caractéristiques	Life Technologies (SOLiD v4)	Illumina (Hi-Seq2000)	Roche 454 (GS FLX Titanium)
Support	Billes + lame	Lame	Billes + plaque de pico-titration (PTP)
Nombre d'échantillons par support	1, 4, 8	8	2, 4, 8, 16
Technique d'amplification	Émulsion PCR	Par pontage sur phase solide	Émulsion PCR
Technique de séquençage	Par ligation	Par synthèse (SBS)	Par synthèse (Pyro-séquençage)
Longueur des lectures (en nucléotides) : fragment / *paired-end*	50 (F3) / 50 (F3) et 35 (F5)	100 / 2 x 100	400 / 2 x 400
multiplexage	4 à 96 échantillons sur une lame		1, 2, 4, 8, 16 (132)
Nombre de *reads*	$0.7x10^9$	$3x10^9$	$0.5x10^9$
Temps de séquençage /échantillon (fragment)	7 jours	8.5 jours	10 heures

TABLE 1.2 – Caractéristiques des trois modèles de séquenceurs les plus répandus

Le modèle GS FLX Titanium de Roche permet ainsi de séquencer des fragments d'ADN longs, de 400 nucléotides, alors que les autres technologies séquencent des fragments courts de 50 à 100 nucléotides. Cependant, le volume de séquences produites (ou « ***reads*** ») est plus limité. Cette technologie est donc largement utilisée pour le séquençage *de-novo* des génomes de grande taille. La longueur de ces *reads* permet un assemblage plus facile du génome d'intérêt bien que la couverture obtenue (*i.e* ; le nombre de fois où une base est séquencée) reste relativement faible.

En revanche, les séquenceurs Illumina et SOLiD™génèrent des *reads* de courte taille en très grand nombre. Cette particularité est très utile pour les applications qui nécessitent un plus grand nombre de lectures par position (c'est-à-dire une grande couverture), telles que l'étude quantitative du transcriptome, la détection spécifique de polymorphismes ou des marques épigénétiques.

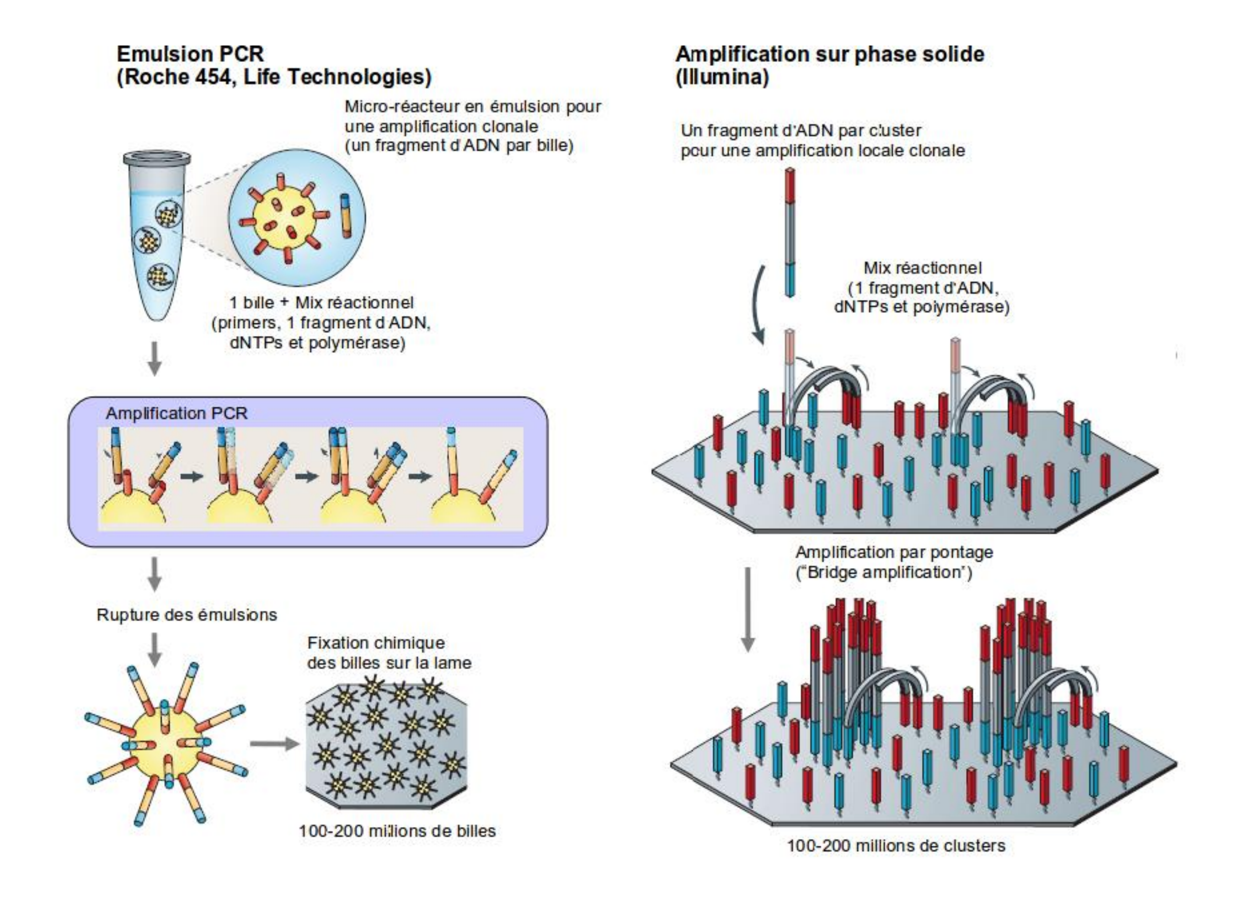


FIGURE 1.7 – Amplification monoclonale des fragments d'ADN pour la constitution de librairies (Adapté de [Metzker, 2010])

Généralités Après réparation des extrémités non cohésives des séquences d'ADN double brins obtenues par fragmentation d'ADN génomique ou d'ADNc, la première étape pour permettre le séquençage consiste en l'ajout d'un **couple de séquences adaptatrices**. Ces adaptateurs permettent la fixation sur bille ou sur lame des fragments d'ADN en vue de leur **amplification par PCR (« *Polymerase Chain Reaction* ») en émulsion** ou par **pontage sur phase solide** (Figure 1.7). Leur séquence ne s'aligne pas sur celle du génome, ce qui permet une amplification PCR spécifique des séquences cibles que l'on souhaite séquencer. Les fragments d'ADN sont ensuite sélectionnés en fonction de leur taille.

On peut également utiliser des adaptateurs particuliers possédant, en plus de la séquence adaptatrice, une courte séquence d'identification spécifique. Cette courte séquence de 5 nucléotides est appelée code barre (« ***barcode*** »). En utilisant un jeu de code-barres unique pour chaque échantillon, cette technique appelée **multiplexage** permet de séquencer plusieurs échantillons dans la même cellule ou dans la même piste. À cette étape, on ne parle plus d'échantillons mais de **librairies**. Ces librairies sont enfin séquencées simultanément lors d'un cycle de séquençage (ou « ***run*** »). Les *reads* obtenus sont ensuite réaffectés automatiquement à chaque échantillon, grâce à l'identification informatique du code barre.

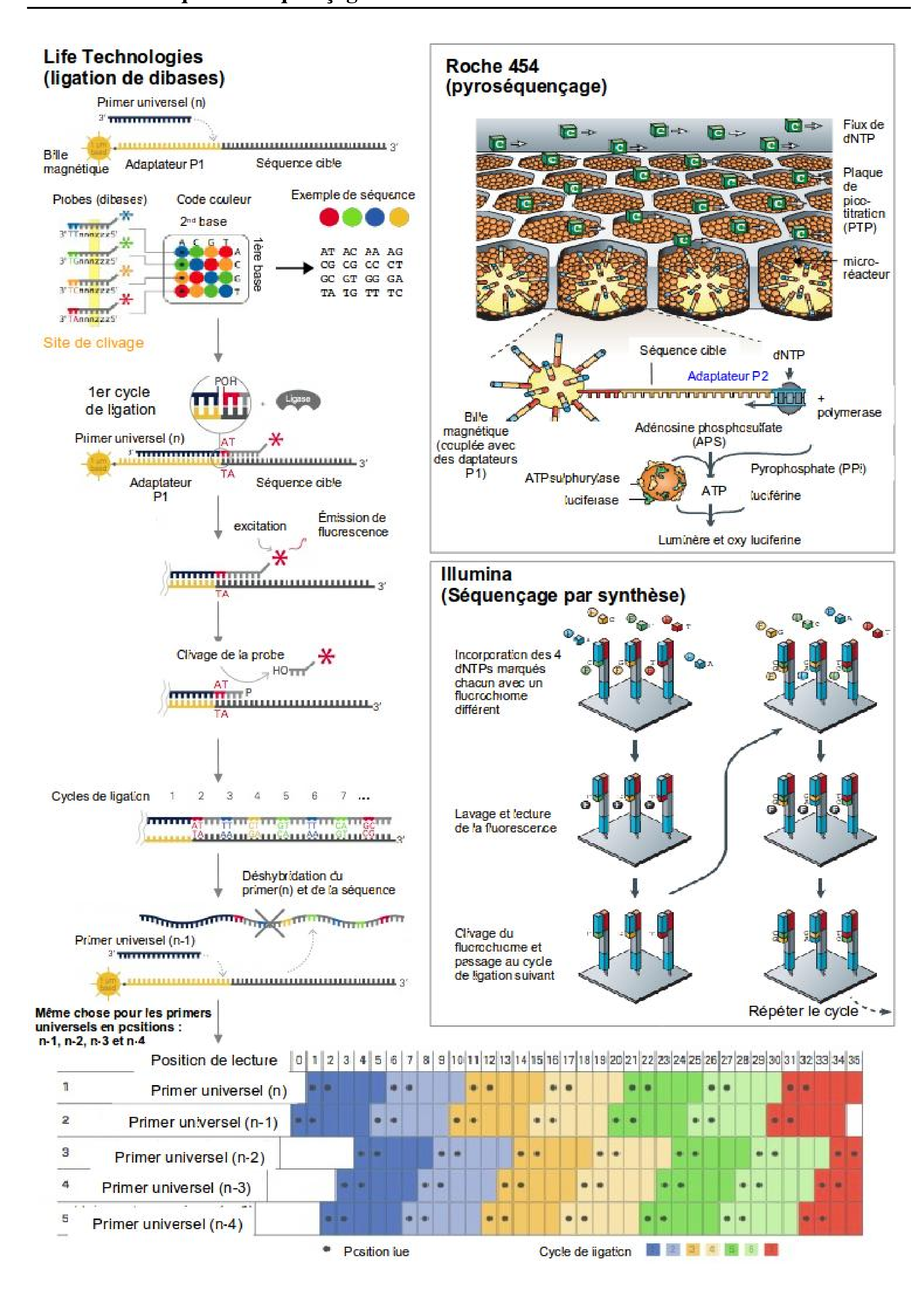


FIGURE 1.8 – Principe des trois technologies majeures de séquençage à très haut débit (Adapté de [Metzker, 2010])

Principe de la chimie du GS FLX Titanium de Roche La technique commercialisée par la société Roche est basée sur l'amplification par **PCR en émulsion** de l'ADN fixé sur des billes, suivie de **pyroséquençage** ([Margulies *et al.*, 2005], [Rothberg & Leamon, 2008]) dans une plaque de pico-titration permettant une lecture en parallèle des millions de fragments d'ADN à séquencer (Figure 1.8). Le modèle le plus puissant actuellement commercialisé est le GS FLX Titanium.

Les librairies, constituées de fragments d'ADN simple brin auxquels sont fixés les adaptateurs, sont mises en présence de billes magnétiques sur lesquelles sont fixées des milliers de copies de la séquence complémentaire de l'adaptateur 1. Une émulsion en concentration limitante d'ADN permet de fixer un seul fragment d'ADN par bille qui sera par la suite amplifié par PCR. Une fois les fragments présents en de multiples copies monoclonales sur les billes, celles-ci sont placées dans des mini-réacteurs, des plaques de pico-titration. Ces plaques permettent une lecture indépendante de chaque bille, donc de chaque séquence par pyroséquençage ([Rougemont *et al.*, 2008], [Droege & Hill, 2008]).

Lors du pyroséquençage, les nucléotides sont rajoutés successivement contrairement aux réactions de séquençage usuelles où les nucléotides sont rajoutés simultanément, chacun étant marqué à l'aide d'un fluorochrome différent. Si le nucléotide présent dans le milieu réactionnel est celui attendu par l'ADN polymérase, il est incorporé dans le brin d'ADN en cours de synthèse libérant ainsi un pyrophosphate (PPi). Ce PPi permet la transformation de l'adénosine 5' phosphosulfate (APS) en adénosine tri-phosphate (ATP) par l'ATPsulfurylase. Cette molécule d'ATP va se coupler à la luciférine pour permettre à la luciférase de transformer la luciférine en oxyluciférine. Cette dernière réaction émet un signal lumineux qui va être capté par le scanner du séquenceur. Les nucléotides en surplus dans le milieu réactionnel sont alors dégradés par une apyrase ce qui permet la lecture de la base suivante ([Ronaghi *et al.*, 1998]). Il est à noter que la taille des séquences lues par cette technique reste faible (inférieure à 400 nucléotides), l'activité enzymatique diminuant au cours du séquençage ([Ahmadian *et al.*, 2006], [Ronaghi, 2001]).

Principe de la chimie du HiSeq2000 d'Illumina La technologie « True-Seq » commercialisée par la société Illumina repose sur l'utilisation d'une lame de verre divisée en 8 pistes linéaires dans lesquelles sont fixées, à haute densité, deux courtes séquences d'ADN de manière aléatoire. Ces deux séquences correspondent aux séquences complémentaires des adaptateurs. À la différence des deux autres technologies qui ont recours à l'amplification par émulsion PCR couplée à l'utilisation de billes magnétiques, les fragments d'ADN possédant les deux adaptateurs vont s'hybrider à la lame de manière homogène (Figure 1.7). L'amplification est ensuite obtenue à l'aide d'amorces spécifiques des adaptateurs. Cette amplification permet la constitution **par pontage** d'un groupe (« *cluster* ») de séquences amplifiées. Le **séquençage par synthèse** (« *Sequencage By Synthesis* » ou SBS, Figure 1.8) de ces *cluster*s va permettre la lecture de l'incorporation d'un fluorochrome spécifique d'une base à chaque ligation, par la prise d'une image à très haute résolution de la lame. Après repérage des *cluster*s, il est possible de reconstituer les séquences ADN pour chacun d'entre eux.

La société Illumina commercialise actuellement plusieurs modèles de séquenceurs ayant la même chimie de séquençage, mais offrant des caractéristiques de séquençage toujours plus compétitives (capacité et configuration). La majorité des données de séquençage à très haut débit étant issue de technologie Illumina (980 séquenceurs / 1670), un grand nombre de programmes d'analyse de données ont initialement été développés pour cette technologie ([Kircher *et al.*, 2011], [Goldfeder *et al.*, 2011], [Kircher *et al.*, 2009]).

Principe de la chimie du SOLiD™de Life Technologies La technologie du SOLiD™(pour « ***Sequencing by Oligonucleotide Ligation and Detection*** ») développée par Life Technologies est basée sur une amplification par **PCR en émulsion** de la même façon que pour le modèle commercialisé par Roche. Les billes sont, dans ce cas, simplement fixées sur une lame de verre. L'originalité de cette technologie repose sur un séquençage de fragments en parallèle et à très haut débit, **par ligation de di-bases couplées à un fluorochrome**. Ainsi, la lecture d'un fluorochrome ne code pas pour une base comme pour la SBS, mais pour la ligation de deux bases successives.

La plateforme IBiSA TGML du laboratoire TAGC Inserm UMR_S 928 étant équipée d'un séquenceur de technologie SOLiD™depuis avril 2009, cette technologie est décrite plus en détail dans ce manuscrit.

Le séquençage du fragment d'ADN est réalisé par l'hybridation d'une amorce complémentaire de la séquence de l'adaptateur P1 (Figure 1.8) et l'ajout de sondes d'une taille de 8 nucléotides couplées à un fluorochrome.

Les bases dégénérées correspondent à un mélange équimolaire des 4 nucléotides par position (Figure 1.9). Les sondes de 8 nucléotides sont complémentaires sur 5 nucléotides, il existe donc 45 sondes possibles soit 1024 sondes au total. L'inclusion de nucléotides universels dans les sondes permet un séquençage efficace et rapide de séquences nucléotidiques plus longues.

Après la ligation d'une sonde, le signal émis par chaque bille est détecté (photographie à haute résolution). Puis les sondes sont clivées en position 5 pour permettre une nouvelle ligation. La lecture des deux premières positions du fragment d'ADN cible est ainsi obtenue. Cette étape est répétée 10 fois, afin de constituer une séquence incomplète de 50 nucléotides (cas du mode fragment et de la lecture de 50 nucléotides). Puis 4 autres cycles de 10 ligations sont réalisés à partir d'amorces s'hybridant également sur l'adaptateur P1 en position n-1, n-2, n-3 et enfin n-4. La combinaison des 5 séquences partielles en code couleur permet de reconstituer la séquence de 50 nucléotides.

1.4.1.2 Le code couleur du SOLiD™ : avantages et inconvénients

Une des particularités du séquençage SOLiD™est que chaque nucléotide est séquencé 2 fois. En effet, ce séquenceur ne lit pas base par base comme les modèles déclinés par Roche ou

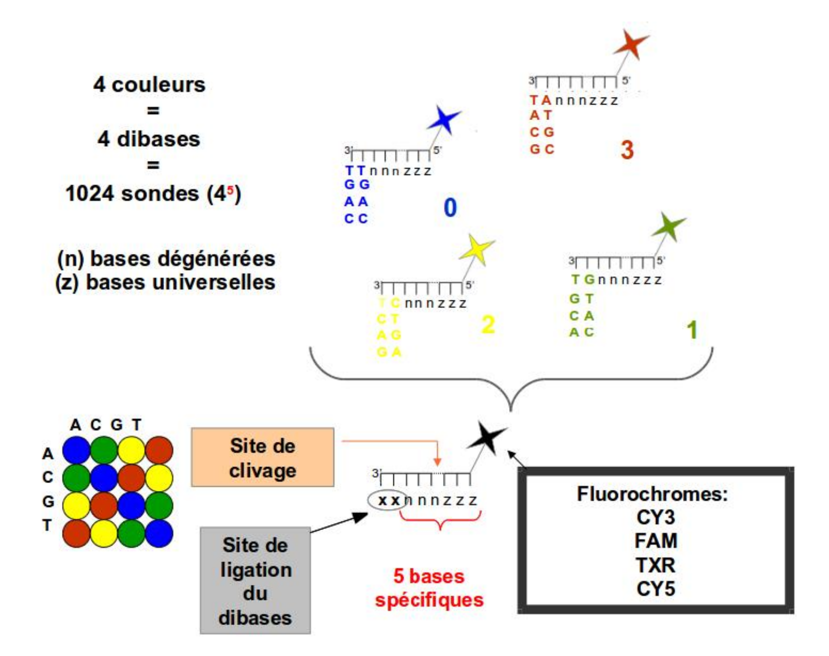
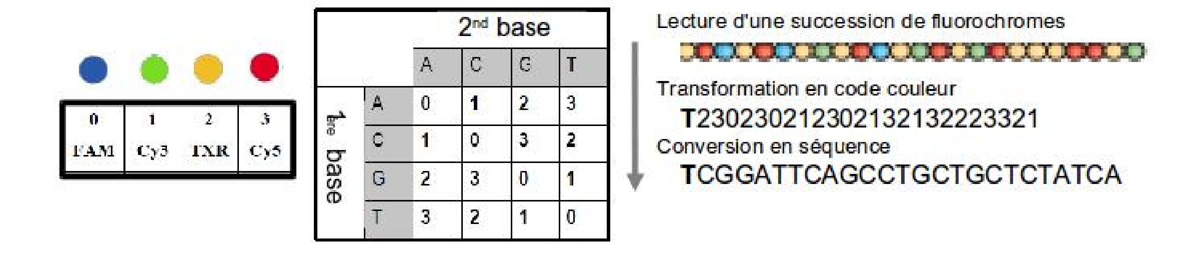


FIGURE 1.9 – Les sondes de la technologie SOLiD™. Chaque sonde de 8 nucléotides est composée de 2 bases complémentaires à la séquence cible (positions 1 et 2) puis de 3 bases dégénérées (n) et enfin de trois bases universelles (z)

Illumina, mais définit la séquence cible par la lecture de di-bases (Figure 1.10). Cette stratégie est particulièrement adaptée à la détection de SNP (« *Single Nucleotide Polymorphisms* ») qui ne diffèrent du génome de référence que par un nucléotide. On peut ainsi aisément distinguer une erreur de séquençage (changement d'une seule couleur), ou « mismatch », d'un véritable SNP qui sera détecté par le changement successif de deux couleurs (Figure 1.11).

Ainsi, à partir de la succession de fluorochromes observés lors du séquençage, pour chaque bille, le SOLiD™génère une séquence en code couleur (au format « csfasta » ; voir partie 5.3.1), dont la première lettre correspond à la dernière base de l'adaptateur P1 (position n ; Figure 1.8).



	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0

FIGURE 1.10 – Conversion des *reads* SOLiD™en séquences nucléotidiques. Chaque couleur code pour un nombre entre 0 et 3 permettant à l'aide de la dernière base de l'adaptateur (dans cet exemple « T ») de reconstituer la séquence génomique

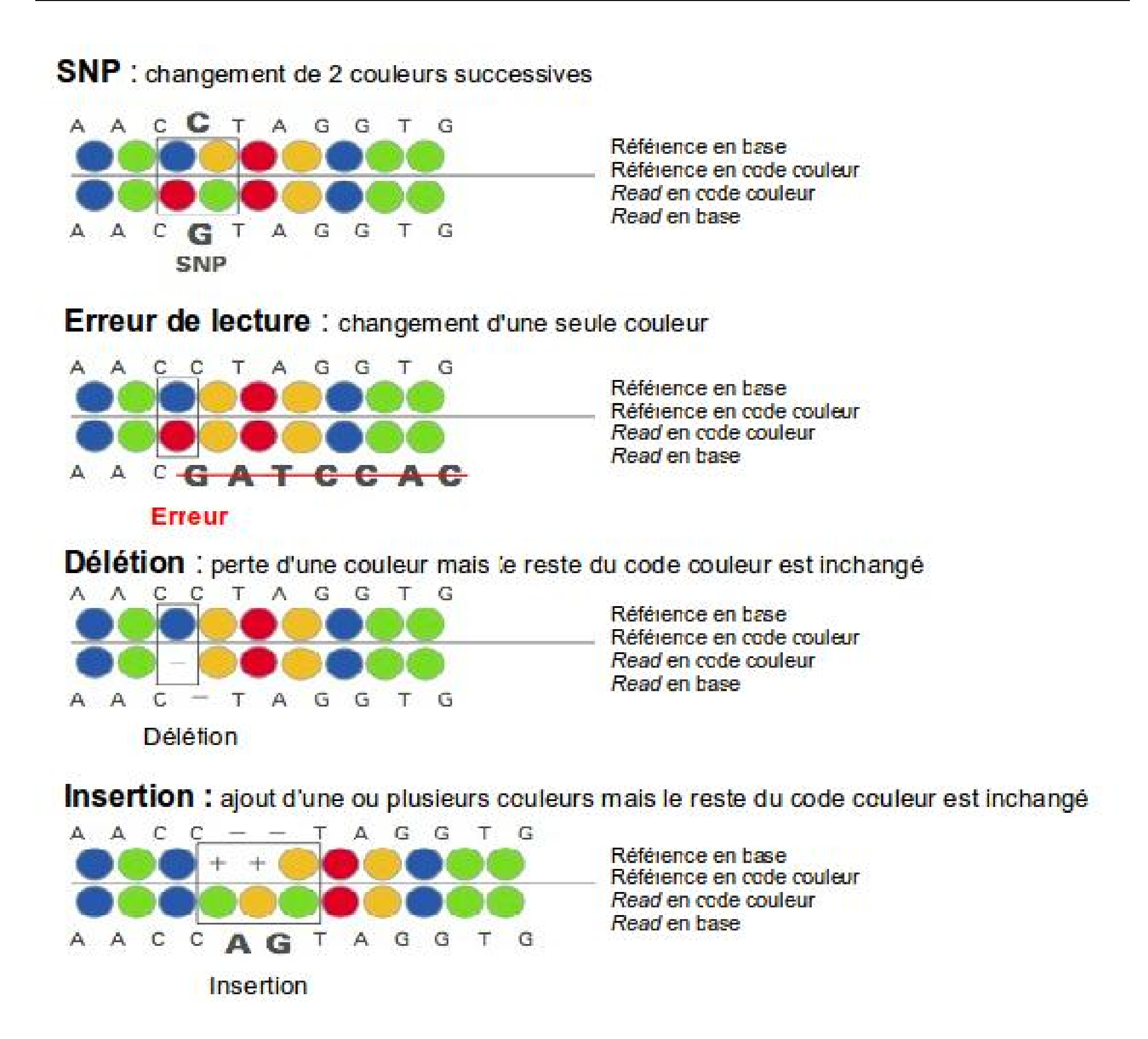


FIGURE 1.11 – Principe de la détection de SNP et small *indels* par la technologie SOLiD™

Il est à noter que ce format présente un inconvénient : dès qu'une erreur de séquençage se produit, le reste de la séquence est erroné (Figure 1.11). C'est pourquoi, afin d'améliorer la qualité des séquences et de visualiser toute erreur, les alignements sont effectués sur la base du code couleur et non directement sur l'interprétation du code couleur en nucléotides.

1.4.1.3 Mode de séquençage

En fonction des projets et du type d'expérience, le mode de séquençage des librairies le plus adapté sera choisi. En effet, il existe 3 modes de séquençage possibles d'un fragment d'ADN : « ***fragment*** », « ***paired-end*** » et « ***mate-pair*** ». Chaque mode requiert des protocoles expérimentaux différents afin de générer les librairies correspondantes. La figure 3.12 (A) représente ces 3 types de librairies dans le cas d'un séquençage de type SOLiD v4.

Excepté pour le mode *mate-pair*, il est également possible de multiplexer les échantillons à l'aide de code-barres rajoutés dans la séquence de l'adaptateur P2.

Fragment Comme le montre la figure 1.12 (A), ce mode de séquençage relativement simple consiste à séquencer, de 5' en 3', 50 nucléotides des fragments d'ADN à partir de l'adaptateur P1. Cela permet le comptage des fragments d'ADN et peut donc être utilisé pour des applications telles que l'étude du transcriptome.

Paired-end De plus en plus utilisé à l'heure actuelle, cette technique permet un meilleur alignement des séquences par le séquençage de 2 fragments d'ADN séparés par 100 à 300 nucléotides. Ainsi, pour le séquenceur de type SOLiD, les fragments d'ADN sont séquencés sur 50 nucléotides de 5' en 3' des fragments d'ADN à partir de l'adaptateur P1 et sur 35 nucléotides de 5' en 3' à partir de l'adaptateur P2 (Figure 1.12 A). La distance approximative entre les fragments F3 et F5 correspond à la taille des fragments d'ADN générés lors de la fragmentation par ultrasons et peut être déterminée par migration sur gel ou par l'utilisation de puces à haute sensibilité du Bioanalyzer (Agilent Technologies). On peut ainsi définir un intervalle de taille de fragment. En effet, il est important d'évaluer ces distances afin de prévoir, lors de l'alignement, l'intégration d'événements de type insertion, délétion, inversion (Figure 1.12 B).

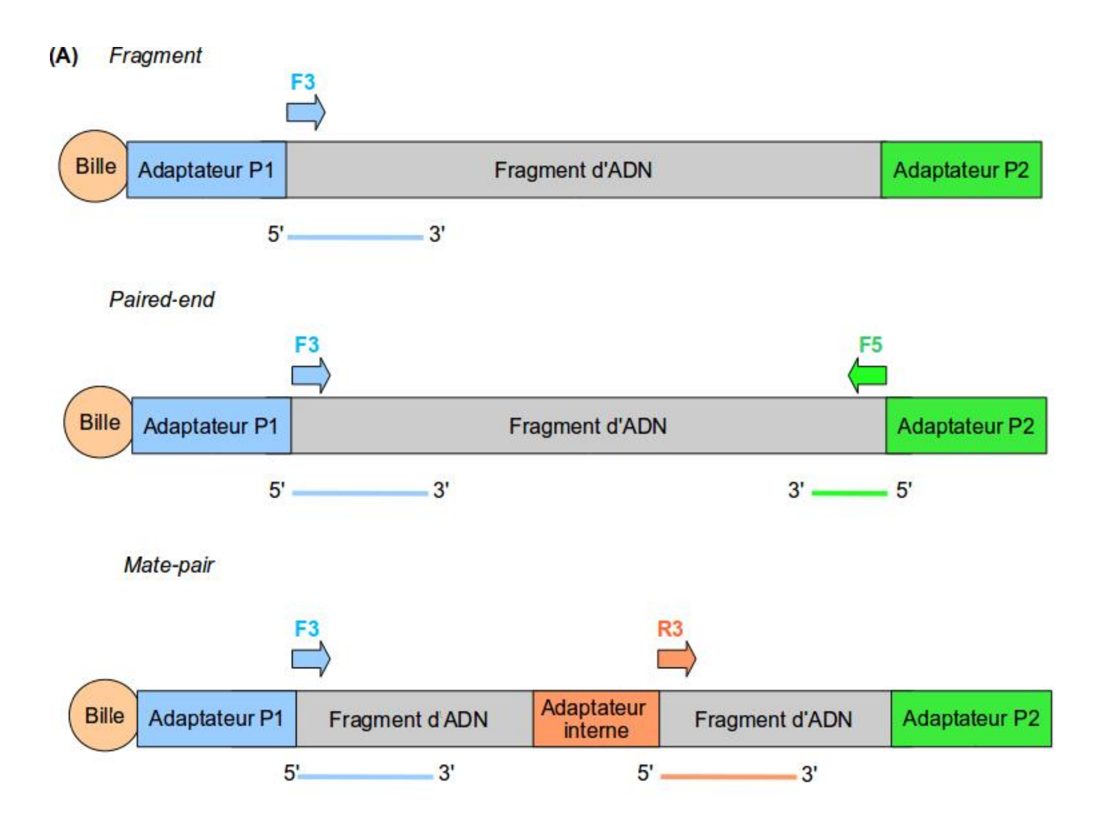
L'utilisation du séquençage en *paired-end* (ou PET, pour « *Paired-End Tag* ») présente divers avantages en fonction du type d'application souhaité (Table 1.3). Ainsi, pour la technique d'Immunoprécipitation de la Chromatine (ChIP, « *Chromatin ImmunoPrecipitation* »), on parle de ChIP-seq quand les librairies sont en fragments et de ChIP-PET ([Wei *et al.*, 2006]) quand celles-ci sont en *paired-end*. Cette technique permet d'accroître la spécificité et la démarcation des sites de fixation des facteurs de transcription.

Comme le montre la table 1.3, ce mode de séquençage est couramment utilisé pour diverses applications puisqu'il permet une nette amélioration de l'efficacité et de la qualité d'alignement des « *reads* ».

Mate-pair Ce mode de séquençage permet de séquencer 2 fragments de même taille (50 nucléotides) et éloignés sur le génome de 1 à 10 kb soit une distance dépassant la taille des fragments nécessaires pour la construction des librairies. Il permet le re-séquençage des génomes (Re-seq) afin d'étudier les remaniements à grande distance ([Shendure *et al.*, 2005]), tels que les « *indels* » (insertions ou délétions), les grandes duplications et délétions, les inversions, les translocations ou encore les anomalies de ploïdie. La construction de librairies en « *mate-pair* » permet ainsi le séquençage orienté de larges fragments d'ADN.

Application	Avantage du PET	Technique et référence
Alignement des *reads*	Augmentation de l'efficacité lors de l'alignement	*Paired-end ditag* (PET) [Ng *et al.*, 2005], [Wei *et al.*, 2006]
	Diminution du coût de séquençage des échantillons	*Paired-end sequencing* (PES) [Holt & Jones, 2008]
	Apport d'informations par rapport aux distances entre les deux fragments séquencés et leur relation (délétion, insertion, inversion)	*Paired-end mapping* (PEM) [Korbel *et al.*, 2007], *Mate-pairs* ([Shendure *et al.*, 2005]), *Paired-end genomic signature tags* (PE-GST) [Dunn *et al.*, 2007]
Transcriptome	Identification des 5' et 3' UTR	*Gene identification signature* (GIS-PET) [Ng *et al.*, 2005]
	Identification de TSS alternatifs	*Gene Scanning CAGE* (GSC-PET) [Carninci *et al.*, 2005]
Epigénétique	Amélioration de la spécificité et de la démarcation des fragments contenant le site d'intérêt	ChIP-PET [Wei *et al.*, 2006]
Variation de la structure du génome	Requis pour le séquençage *de novo*	DNA-PET [Hillmer *et al.*, 2011]

Table 1.3 – Applications et avantages du *paired-end* pour les techniques de séquençage à très haut débit



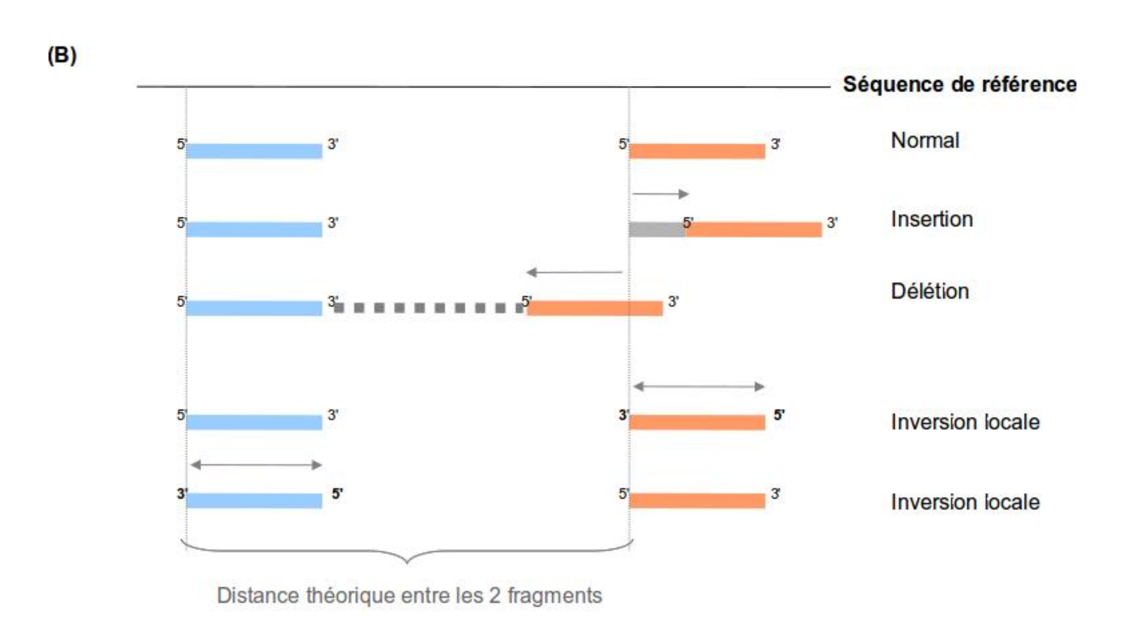


FIGURE 1.12 – (A) Les différents modes de séquençage : « fragment », « *paired-end* » et « *mate-pair* » (B) Apport du mode « *paired-end* » pour la détection d'évènements d'insertion, de délétion ou d'inversion

1.4.1.4 Les nouvelles générations de séquenceurs à haut débit

Avec les développements technologiques permanents est apparue la quatrième génération de techniques de séquençage ([Glenn, 2011]). Elle regroupe les nouvelles technologies telles que **PacBio RS** (Pacific Bioscience Inc.) et **Ion Torrent** (Life Technologies). Basées sur le séquençage par synthèse, elles utilisent deux nouvelles chimies.

PacBio RS repose sur l'analyse de molécule unique par des **réactions de séquençage en temps réel (« *Single Molecule Real Time* » ou SMRT^{TM})** (Figure 1.13). Elle utilise le processus très efficace et précis de réplication de l'ADN par l'ADN polymérase. Cette enzyme, fixée au fond des puits, se lie à un unique fragment d'ADN à séquencer. Cependant, elle ne peut intégrer que quelques nucléotides marqués avant de s'arrêter à cause de l'encombrement stérique de ces nucléotides. Pour y remédier, la méthode SMRT^{TM}utilise un nucléotide marqué non pas au niveau de la base mais au niveau de la chaîne phosphate. Le système de détection enregistre un film chronologique de ces événements à la différence des systèmes antérieurs basés sur l'analyse d'images à très haute résolution.

Quant à la technologie Ion Torrent, elle est basée sur des **puces semi-conductrices** constituées de puits (Figure 1.13) ([Rothberg *et al.*, 2011]). Elle suit le principe, publié en 1968, selon lequel un proton est relargué dès lors qu'un nucléotide est incorporé dans l'ADN par la polymérase ([Narurkar *et al.*, 1968]). Il en résulte un **changement de pH local**, mesuré par un détecteur sensible à la variation d'ions. Cette dernière technologie ne nécessite ni caméra, ni scanner, ni cascade enzymatique, ni fluorochrome ou chimioluminescence. Elle diffère en cela de toutes les autres basées sur la détection d'un signal lumineux, d'où la dénomination anglaise de « *post light sequencing technology* ».

De plus, avec la course à l'amélioration des technologies de séquençage, les principaux fournisseurs ont eux aussi développé ou acquis de nouveaux appareils de paillasse pouvant séquencer très rapidement des librairies avec néanmoins un débit moins important : **MiSeq** (Illumina), **Ion Torrent** (acquis en octobre 2010 par Life Technologies), **GS Junior** System (Roche).

1.4.2 Techniques d'analyses basées sur le séquençage HTS

La technologie de séquençage à très haut débit offre de nombreuses applications allant de l'étude de l'épigénétique au transcriptome en passant par la génomique (Figure 1.14 et Table 1.4).

Cependant, le coût d'utilisation du séquençage à très haut débit reste particulièrement élevé, ce qui explique que, pour l'étude du transcriptome, l'utilisation de puces à ADN pangénomiques décrites précédemment est encore très répandue. Cependant, l'utilisation du RNA-seq par « *Whole Transcriptome Shotgun Sequencing* » (WTSS) ou « *Serial Analysis of Gene Expression* » (SAGE-seq) est utile pour l'étude globale des transcrits (mRNA, snRNA, lincRNA, miRNA), la détection de transcrits alternatifs et de nouveaux gènes.

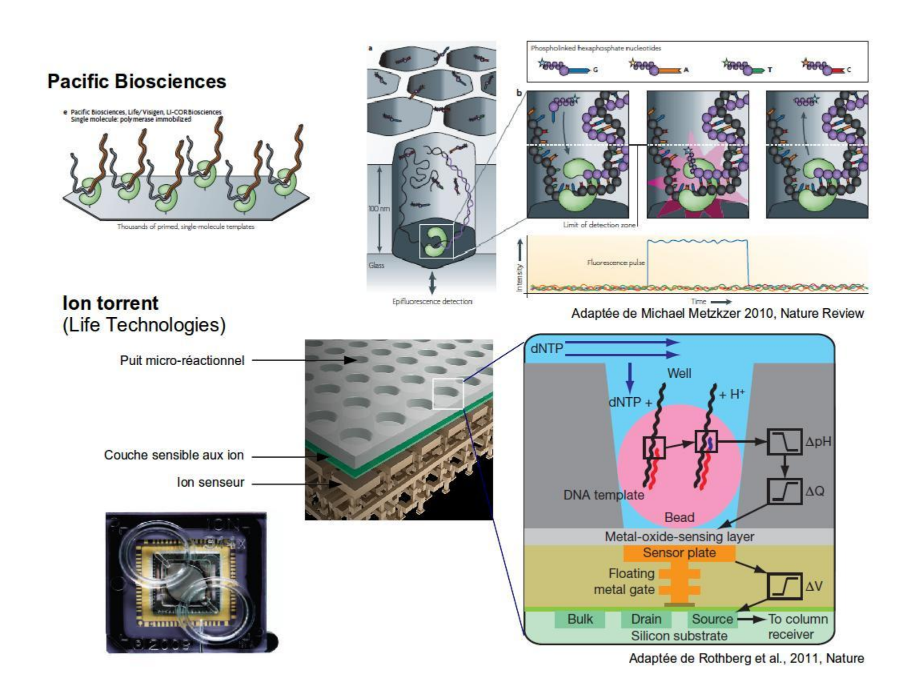


FIGURE 1.13 – La nouvelle génération de séquenceurs. (A) La technologie de Pacific Biosciences sur le principe de SMRT™(B) L'Ion Torrent et sa puce semi-conductrice pour la lecture d'un différentiel de pH (Adapté de [Metzker, 2010] et [Rothberg *et al.*, 2011])

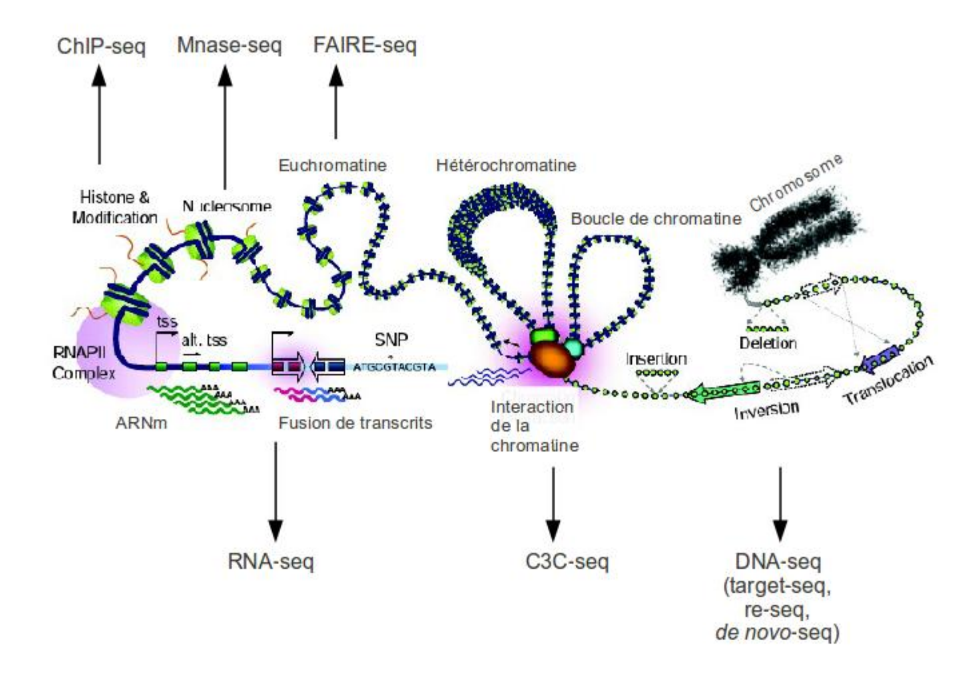


FIGURE 1.14 – Les différentes études rendues possibles par le HTS avec divers niveaux d'abstraction (adapté de [Fullwood *et al.*, 2009])

L'utilisation du séquençage à très haut débit nécessite encore de nombreuses mises au point expérimentales et bioinformatiques pour une analyse performante et aboutie des résultats générés. En effet, les données produites représentent plusieurs Gigabytes (Gb) par échantillon, dont l'exploitation requiert de puissants ordinateurs de calcul, et dont l'archivage nécessite de grandes capacités de stockage.

Le choix de la technologie de séquençage sera opéré en fonction des applications souhaitées. Ainsi, le modèle de Roche sera choisi pour le séquençage *de-novo*, en raison de la taille plus importante des *reads* ce qui facilite l'assemblage des génomes. Les technologies de type SOLiD™et Illumina seront préférées pour les études épigénétiques et la détection de polymorphismes tels que les SNP, les insertions et délétions.

1.4.2.1 Etude de la régulation épigénétique

Le contrôle de la structure dynamique de la chromatine est une composante essentielle de la régulation transcriptionnelle dans les cellules eucaryotes. Diverses techniques à très haut débit ont été mises au point pour l'étude de la régulation épigénétique du génome à différents niveaux : la conformation des chromosomes (3C-seq), l'ouverture de la chromatine (FAIRE-seq ou traitement à la DNAse I), la position des nucléosomes (MNase-seq), les modifications des histones et la liaison des facteurs de transcription (ChIP-seq) ou bien encore la méthylation de l'ADN (methyl-seq). Il est à noter que, pour la technique du ChIP-seq, une liaison covalente des protéines à l'ADN (« *cross-link* ») réalisée par un traitement de fixation au formaldéhyde est requise, afin de pouvoir cartographier les sites de liaison des facteurs de transcription (voir partie 5.1.2).

L'expression d'un gène peut être contrôlée par l'interaction directe de son promoteur avec des éléments de régulation localisés à une longue distance sur le chromosome, ou dans de rares cas sur d'autres chromosomes. La technique du **3C-seq** permet ainsi la Capture de la Conformation des Chromosomes (« *Capture Chromosome Conformation* » ou 3C). Elle a été développée pour l'analyse de la chromatine à une échelle supérieure. Les régions du génome bien qu'éloignées peuvent alors être juxtaposées par bouclage de la chromatine et deviennent ainsi contiguës, à condition que la chromatine soit ouverte.

En effet, l'ouverture de la chromatine permet la liaison de protéines régulatrices à l'ADN. Elle peut être étudiée par la technique appelée couramment « *Formaldehyde-Assisted Isolation of Regulatory Elements* » (FAIRE) ([Giresi *et al.*, 2007], [Song *et al.*, 2011], [Nammo *et al.*, 2011]). Le **FAIRE-seq** permet de cartographier certaines régions ouvertes de la chromatine, et ainsi de définir des régions régulatrices. En effet, seul 1 à 2% du génome est composé de régions ouvertes de la chromatine dans un type cellulaire donné et dans des conditions particulières ([Song *et al.*, 2011]). On obtient alors des informations sur des régions régulatrices sur lesquelles des facteurs de transcription peuvent venir se lier à l'ADN ou bien encore des sites pour lesquels les histones nucléosomales subissent des modifications post-transcriptionnelles.

Domaine d'étude	Technique	Description
Epigénétique	**ChIP-seq**	**Cartographie des sites de fixation de facteurs de transcription et des modifications d'histones [O'Geen *et al.*, 2011]**
	Methyl-seq	Cartographie des sites de méthylation de l'ADN (ilôts CpG) [Li *et al.*, 2011] [Wu *et al.*, 2011*a*]
	3C-seq	Recherche des interactions chromatiniennes à longue distance : « *Capture Chromosome Conformation* » [Splinter *et al.*, 2004] [Gheldof *et al.*, 2012]
	Mnase-seq	Position des nucléosomes
	FAIRE-seq	Cartographie des régions régulatrices [Giresi *et al.*, 2007]
Transcriptome	RNA-seq	Ou « Whole Transciptome Shotgun Sequencing » (WTSS) permet le séquençage de tous les ARNs pour : – l'etude du transcriptome ; – la découverte de nouveaux gènes ; – l'étude des sites d'épissage (épissage alternatif). [Hong *et al.*, 2011] [Clark *et al.*, 2011] [Bainbridge *et al.*, 2011]
	SAGE-seq	« *Serial Analysis of Gene Expression* » (SAGE) ou « *Digital Gene Expression* » (DGE) pour le comptage relatif de transcrits
	sRNA-seq	Identification des petits ARNs (*miRNA, lincRNA, snRNA, ncRNA*) [Gommans & Berezikov, 2012]
	GRO-seq	« *Global run-on sequencing* » [Core *et al.*, 2008]
Génomique	De novo-seq	Séquençage d'un génome, métagénomique
	Re-seq	Re-séquençage des génomes pour l'identification des variations inter-individuelles (SNP, *indels* (insertions/délétions), grandes duplications et délétions, inversions, translocations, anomalies de ploïdie, etc.) [Pareek *et al.*, 2011]
	Target-seq	Le séquençage ciblé (e.g. gène, région chromosomique sur plus de 20Mb ou exome complet) d'une collection d'échantillons (à partir de puces à ADN) pour la détection de polymorphismes (SNP, *indels*)

Table 1.4 – Les principales applications du séquençage à très haut débit. En gras figure l'application qui sera développée plus en détail dans le chapitre 5 de ce manuscrit

Ainsi, l'identification *in vivo* par ChIP de sites de fixation de facteurs de transcription et de modifications N terminales d'histones est maintenant possible à très haut débit par le séquençage des fragments immunoprécipités (**ChIP-seq**). Cette méthode permet une meilleure résolution des sites potentiels de liaison des facteurs de transcription que la technique antérieure utilisant l'hybridation des fragments d'ADN immunoprécipités sur des puces à ADN ou à oligonucléotides pangénomiques (désignée par l'acronyme anglais ChIP-on-chip, voir partie 5.1.1). **La méthode du ChIP-seq sera décrite en détail dans la partie suivante de ce manuscrit, car le traitement des données ainsi obtenues a nécessité le développement d'un pipeline d'analyse spécifique sur la plateforme TGML ce qui a constitué une partie de mon travail de thèse, présenté dans le Chapitre 5.**

En plus des modifications des histones, le positionnement de nucléosomes peut être étudié par une digestion de la chromatine à la DNAseI ou à la **MNase** (« Micrococcal Nuclease » ou « S7 Nuclease ») en présence de cations bivalents. Cette endonucléase effectue des coupures doubles brins de l'ADN entre les nucléosomes. On obtient ainsi des fragments d'ADN de 146 nucléotides environ, soit la taille du fragment d'ADN enroulé autour d'un nucléosome. Par séquençage, il est possible de déterminer les positions de ces nucléosomes sur le génome.

Enfin, l'étude de la méthylation de l'ADN par les techniques appelées **methyl-seq** et « *Reduced Representation Bisulfite Sequencing* » (RRBS), permet la cartographie et la quantification du niveau de méthylation des cystosines (methyl-C) au niveau des îlots ou des dinucléotides CpG sur l'ensemble du génome ([Wu *et al.*, 2011*a*], [Hansen *et al.*, 2011], [Lan *et al.*, 2011]).

1.4.2.2 Etude du transcriptome

Le HTS peut également être utilisé pour l'étude du transcriptome. Les ARN d'intérêt sont rétrotranscrits en ADNc puis chaque ADNc est séquencé. Ces données procurent des informations sur le contenu (qualitatif et quantitatif) en ARN des échantillons. Étant plus sensible que les puces à ADN, le séquençage permet une réelle quantification des transcrits sans effets de saturation du signal sur une plus grande échelle dynamique. Cette méthode peut être utilisée dans le cadre de diverses applications :

- l'identification de nouveaux gènes (transcriptome *de-novo*) ;
- l'identification de régions transcrites mais non traduites (UTRs), de régions charnières intron/exon, de transcrits alternatifs (par épissage alternatif), de codons start ;
- l'identification d'unités non codantes, incluant les ARN non codants, les micro-ARN précurseurs et les autres ARNs non traduits ;
- la détermination du niveau de transcription des gènes.

En HTS, le transcriptome est étudié principalement par deux techniques :

- le **RNA-seq** (« *Whole Transcriptome Shotgun Sequencing* », WTSS) est un outil pour la transcriptomique permettant le séquençage de tous les transcrits d'un échantillon ;
- le **SAGE-seq** (« **Serial Analysis of Gene Expression** ») également appelé DGE-seq pour « *Digital Gene Expression* », précédemment utilisé pour le séquençage des EST

grâce au clonage en série de fragments très courts d'ADNc dans un vecteur plasmidique ([Velculescu *et al.*, 1995]). Cette application permet l'analyse du niveau d'expression d'un grand nombre de gènes via l'identification de séquences en 5'UTR, appelées « étiquettes » (ou « *tags* ») et leur comptage.

L'analyse du transcriptome par séquençage à très haut débit est rapidement devenue un atout précieux pour l'étude de maladies telles que le cancer ([Morin *et al.*, 2008]). Cette approche permet l'acquisition de données d'expression de gènes à l'échelle pangénomique, de la même façon que les approches antérieures utilisant les puces à ADN. Mais elle présente des avantages sur les puces à ADN : elle permet notamment d'obtenir en une seule fois beaucoup plus d'informations, telles que les fusions de gènes, les transcrits alternatifs, les mutations post-transcriptionnelles ou encore l'étude des ARNs non codants (miRNA, lincRNA) ([Linsen *et al.*, 2009]). De plus, la détection des transcrits n'étant pas limitée par les sondes fixées sur le support (cas des puces), on peut identifier de nouveaux transcrits jamais observés auparavant (variants d'épissage, lincRNA).

En fonction de l'application, on choisira l'utilisation des modes fragment ou *paired-end* : la stratégie de séquençage fragment étant préférée pour le « *tag counting* », alors que le *paired-end* permet l'identification de fusion de transcrits, d'épissages alternatifs...

Il est à noter qu'il existe maintenant des kits de séquençage de l'exome et de re-séquençage d'une région ciblée du génome : sur puces à ADN de capture ou sur billes magnétiques ([Clark *et al.*, 2011]). Ces techniques proposent des sondes chevauchantes de taille fixe (60 nucléotides décalés d'un pas de 3 nucléotides, pour Agilent Technologies) couvrant les régions d'ADN à étudier pouvant aller jusqu'à 10Mb. Il est important d'utiliser une séquence ne contenant pas d'éléments répétés ; pour cela, on peut « masquée » la séquence à l'aide du logiciel RepeatMasker ; ceci est généralement intégré dans les logiciels de conception de puces à capture. Les fragments d'ADN ou d'ADNc s'hybridant aux puces ou aux billes seront déshybridés puis séquencés.

1.4.2.3 Autres types d'applications

Il existe diverses techniques permettant l'acquisition de données sur un génome : le séquençage *de-novo* (*de-novo*-seq), le re-séquençage (re-seq) ou bien encore le séquençage de régions ciblées (target-seq). Elles ne nécessitent aucun traitement des cellules. L'ADN génomique est ainsi extrait et séquencé.

La séquence *de-novo* d'un génome peut être obtenue avec une couverture suffisante de manière beaucoup plus rapide et moins coûteuse qu'avec les précédentes technologies. Il n'est plus nécessaire de réaliser des clonages bactériens et autres expériences pour amplifier les fragments d'ADN. Désormais tous les fragments d'ADN sont séquencés. Il est préférable d'utiliser la technologie Roche. En effet, elle génère des fragments longs qui facilitent de reconstituer la séquence génomique (par le biais de scaffolds) ; celle-ci contenant généralement des séquences répétées, elles seraient difficiles à insérer correctement dans le génome sans ce type de séquençage. Il existe également des approches hybrides utilisant la technologie

Roche pour générer des scaffolds et celle d'Illumina ou de Life Technologies permettant d'améliorer la couverture. De plus, de récentes approches mêlent des données génomiques et transcriptomiques pour reconstruire le génome en s'aidant des séquences des ARNc obtenus par HTS.

Pour les génomes dont la séquence est connue, la technique appliquée est le **re-seq**, c'est-à-dire le re-séquençage partiel ou entier du génome. Cette approche permet non seulement l'étude des variations interindividuelles telles que les SNPs, les petits « *indels* » (insertions/délétions), mais également celle des grandes duplications et délétions, des inversions, des translocations, ou encore des anomalies de ploïdie (CNV pour « *Copy Number Variation* » en anglais). Elle est très utilisée pour l'étude des remaniements chromosomiques dans les cancers.

Toutefois, comme pour le RNA-seq, il est parfois nécessaire de sélectionner une ou plusieurs régions génomiques afin qu'elles soient enrichies lors du séquençage. Cette capture des fragments d'ADN localisés au sein d'une région précise du génome est appelée target-seq. Le **target-seq** permet l'analyse ciblée de régions candidates provenant de l'étude de liaisons génétiques, afin d'identifier de nouveaux SNPs et/ou *indels* associés à une maladie ou un phénotype particulier.

Le séquençage permet également de détecter des génomes viraux ou bactériens intégrés au génome de leur hôte après l'avoir infecté. Cette approche de métagénomique vise à étudier directement des organismes microbiens dans leur environnement sans passer par une étape de culture en laboratoire. En conclusion, l'utilisation du séquençage à très haut débit offre de nombreux avantages :

- le multiplexage des échantillons ;
- l'utilisation d'amorces spécifiques des adaptateurs pour réaliser les amplifications PCR ;
- l'automatisation des étapes de préparation des librairies ;
- la réduction de biais d'amplification grâce à la suppression d'étapes de clonage bactérien, de purification sur gel ;
- la quantification est rendue possible, car chaque molécule est séquencée ;
- la couverture de séquençage est élevée (selon le multiplexage) ;
- les applications sont très variées ;
- ces techniques sont plus sensibles que les approches antérieures ;
- le re-séquençage ou le séquençage *de-novo* avec une couverture de séquençage élevée.

1.5 Apports des techniques de puces à ADN et de séquençage à très haut débit

Avec le développement des techniques à haut débit comme les puces à ADN puis le séquençage à très haut débit, une meilleure caractérisation des altérations au niveau transcriptionnel et de la régulation de l'expression des gènes a pu être étudiée. Ceci a conduit à la classification de cancers en groupes selon l'expression de quelques gènes devenus de bons indicateurs de la progression ou du type de tumeur.

Golub et collaborateurs ont ainsi pu proposer en 1999 une signature moléculaire portant sur l'analyse de profils d'expression de différentes leucémies : des leucémies lymphoïdes aiguës (ALL pour « *Acute Lymphoblastic Leukemia* ») et des leucémies myéloïdes aiguës (AML pour « *Acute Myeloid Leukemia* ») ([Golub *et al.*, 1999]). De nouvelles approches de classification et de sélection de gènes ont été mises au point et testées à partir de ce jeu de données ([Inza *et al.*, 2004], [Wu *et al.*, 2011*b*], [Wang & Simon, 2011], [Moorthy & Mohamad, 2011]).

Le même type de classification a été réalisé plus tard par Sorlie et collaborateurs en 2003, puis par Bertucci et collaborateurs en 2004 sur des échantillons de cancers du sein ([Sorlie *et al.*, 2003], [Bertucci *et al.*, 2004]). Ces études, ainsi qu'une étude histologique de ces mêmes échantillons, sont à la base de la classification des tumeurs du sein en 5 groupes : basale, luminale A, luminale B, ERBB2+ et normale. Elles ont permis la définition de biomarqueurs, comme le gène ERBB2 (pour « v-erb-b2 erythroblastic leukemia viral oncogene homolog 2 »), permettant de caractériser les tumeurs et d'ajuster ainsi le traitement en fonction de leurs caractéristiques ([Bertos & Park, 2011]).

De même que les puces à ADN, le séquençage à très haut débit est de plus en plus utilisé. De nombreuses publications paraissent pour permettre l'analyse de pathologies au niveau épigénétique par ChIP-seq, et au niveau transcriptionnel par RNA-seq. Ainsi, les techniques HTS ont pu être utilisées pour étudier par exemple (1) les cibles de facteurs de transcription pour différents types de cancers, comme par exemple FOXA1, ER et CTCF pour des lignées de cancers du sein ([Ross-Innes *et al.*, 2011], [Hurtado *et al.*, 2011]), (2) les exomes qui diffèrent énormément entre les lignées cellulaires ([Chang *et al.*, 2011]), avec en particulier (3) les ARN non codants et leurs implications dans le cancer ([Martens-Uzunova *et al.*, 2011], [Ferdin *et al.*, 2010]). Ceci nous permet d'en apprendre d'avantage sur le génome, notamment dans le cas de cancers ([Meyerson *et al.*, 2010], [Aburatani, 2011]). Même si les principales études portent sur le cancer, il existe des études plus fondamentales ou touchant à d'autres pathologies telles que des infections bactériennes ou virales ([Olguin-Lamas *et al.*, 2011], [Lu *et al.*, 2010]).

Ces techniques ont rendu des projets de grande envergure possibles : « *Therapeutically Applicable Research to Generate Effective Treatments* » (TARGET) a pour une analyse d'intégration de plusieurs techniques allant du transcriptome au re-séquençage en passant par la détection de mutations dans le cas de plusieurs cancers touchant les enfants, ceci afin de sélectionner de nouvelles molécules thérapeutiques. Quant au projet « *European Prospective Investigation into Cancer and Nutrition* » (EPIC), il porte non seulement sur la détermination de signatures moléculaires de cancers, mais aussi sur l'impact de facteurs tels que le tabac, la nutrition.

Le développement du séquençage à très haut débit a également permis la création du projet 1000 génomes en 2008 qui a pour ambition de caractériser les variations génomiques par le séquençage d'une large collection de génomes, puis du projet 1000 épigénomes en 2010 par l'« *International Human Epigenome Consortium* » (IHEC), qui a pour objectif la caractérisation

FIGURE 1.15 – Diagramme des objectifs du consortium travaillant sur le décryptage des épigénomes humains : le IHEC (International Human Epigenome Consortium). Cette figure est issue du site internet du consortium IHEC.

d'au moins 1000 épigénomes (1 par tissu du corps humain), incluant les modifications d'histones, les positions des variants d'histones, le remodelage des nucléosome, la méthylation de l'ADN, l'étude des ARN non codants (Figure 1.15). Plus récemment, le projet BLUEPRINT5 a pour objectif l'établissement des épigénomes des différentes lignées cellulaires hématopoïétiques saines et cancéreuses. Ces études sont très importantes car elles permettent de connaître l'épigénome normal de tissus et de pouvoir ainsi les comparer à ceux de pathologies, en particulier dans le cas de maladies touchant spécifiquement un ou plusieurs tissus donnés, comme le coeur, le cerveau, le foie ou bien encore un lignée donnée de cellules immunitaires.

1.6 Langages de programmation pour l'analyse de données

Comme cela sera présenté et discuté ultérieurement, l'analyse de données issues des puces à ADN et du séquençage à très haut débit nécessite des développements bioinformatiques constants : des programmes avec ou sans interface graphique, des services web ou bien encore des bases de données. Ceci est rendu possible par l'utilisation de techniques de programmation adaptées au besoin des biologistes et des bioinformaticiens.

En effet, la bioinformatique peut être définie très simplement comme l'analyse automatique et programmatique des données biologiques dans le but d'en extraire de l'information. Cette discipline constitue la « biologie *in silico* », par analogie avec *in vitro* ou *in vivo*. Elle est un champ de recherche multidisciplinaire où travaillent de concert biologistes, médecins,

informaticiens, mathématiciens, physiciens et bioinformaticiens, dans le but de résoudre un problème scientifique posé par la biologie. Ce terme peut également décrire (par abus de langage) toutes les applications informatiques résultant de ces recherches.

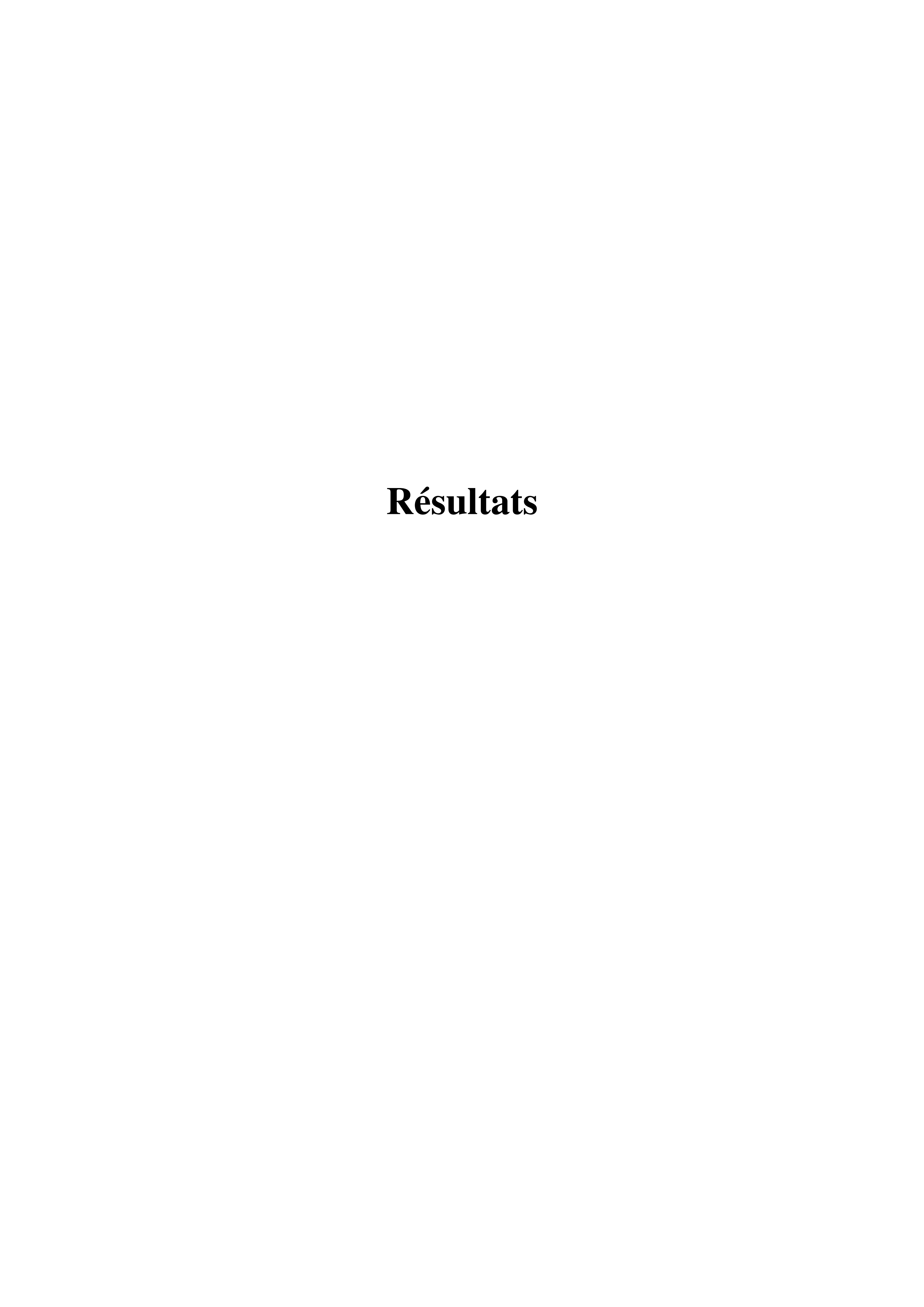
Le développement de ces applications a nécessité la création de nombreux langages, en fonction des besoins des programmateurs ; ceux-ci sont regroupés en styles ou paradigmes :

- **les langages procéduraux** sont des langages où une procédure, appelée également fonction, correspond à une suite d'instructions devant être effectuées sans erreur dans un ordre précis. On distingue parfois procédure et fonction par la caractéristique qu'une procédure ne renvoie pas de résultat ;
- **les langages à objets** utilisent des objets qui correspondent à des structures sémantiques indépendantes et rassemblent également des données et des traitements ;
- **les langages de requête** sont destinés à interroger et manipuler les bases de données ;
- **les langages de définition** de données ne permettent pas de traiter des données mais uniquement de décrire leur structure sous forme de listes ou d'arborescences et les instances de ces structures.

Cependant, un langage peut être associé à plusieurs paradigmes. La table 1.5 ci-après regroupe les différents langages utilisés pendant ma thèse.

Langage	Procédural	Objet	Requête	Définition
bash	x			
C	x			
gawk	x			
java	x	x		
LaTeX				x
perl	x	x		
R	x	x		
SQL	x		x	
XML				x

TABLE 1.5 – Paradigmes associés aux langages utilisés durant cette thèse



Résultats

Chapitre 2

Contrôle qualité et normalisation de données de puces à ADN

Sommaire

Le but d'une expérience de puce à ADN est d'identifier les transcrits dont le niveau d'expression varie entre différentes conditions biologiques d'intérêt. Cependant, ces sources de variation peuvent également être dues, en partie voire en quasi-totalité, à des biais expérimentaux. Afin d'analyser au mieux les données de puces à ADN mais surtout de pouvoir les comparer entre elles lors de l'analyse de données (voir Chapitre 3), il est important de prendre grand soin lors de toutes les étapes expérimentales, de pré-traitement et de normalisation des données brutes afin de limiter ces biais et/ou de les corriger.

C'est pourquoi au laboratoire TAGC Inserm UMR_S 928, nous avons développé notre propre outil permettant l'analyse des données brutes de puces Agilent™réalisées au niveau de la plateforme TGML.

2.1 Obtention des données brutes d'expression

2.1.1 Conception du plan d'expérience et biais techniques

La conception du plan d'expérience est la première étape importante pour obtenir des données de qualité. En effet, il est très important de bien concevoir l'expérience car les sources de variabilité expérimentale sont nombreuses ([Mutter *et al.*, 2004], [Ransohoff & Gourlay, 2010]). L'une des premières sources de variabilité est directement

liée au matériel biologique lui-même bien souvent hétérogène. Cela est particulièrement vrai dans le cas des tumeurs qui sont constituées de nombreuses populations cellulaires très différentes dans la plupart des cas. De plus, le prélèvement des cellules cancéreuses par biopsie s'accompagne souvent de celui de cellules saines. Pour des expériences de puces à ADN, l'idéal serait de travailler sur des populations de cellules homogènes et dont le cycle de division cellulaire serait synchronisé, ce qui n'est pas réalisable dans le cas des biopsies.

Lors de la conception d'une expérience destinée à l'étude d'expression par puces à ADN Agilent^{TM}, il faut avant tout choisir l'approche à utiliser : ***one-color*** ou ***two-colors***, en fonction du nombre d'échantillons et de la question biologique posée ([Oberthuer *et al.*, 2010]). Avec l'approche *one-color*, un seul échantillon marqué à la cyanine-3 (Cy3) est hybridé sur la puce, alors que l'approche *two-colors*, permet l'hybridation simultanée sur la même puce de deux échantillons marqués par des fluorochromes différents : Cy3 et cyanine-5 (Cy5), généralement. Le principal intérêt des puces *two-colors* est de pouvoir comparer directement deux échantillons hybridés sur la même puce en réduisant ainsi les biais techniques inhérents à l'utilisation de puces. Cette stratégie *two-colors* permet non seulement de s'affranchir de certaines variabilités techniques mais aussi d'augmenter la sensibilité et la précision dans la détermination des niveaux d'expression différentiels entre des paires d'échantillons. Cependant, l'efficacité d'incorporation des nucléotides fluorescents varie en fonction du fluorochrome utilisé : les nucléotides marqués par Cy5 sont moins efficacement incorporés que ceux porteurs de Cy3 en raison de l'encombrement stérique des différents nucléotides marqués (dUTP-Cy5 et dUTP-Cy3), ce qui induit des variations d'intensité non imputables à une expression différentielle des gènes. Cette approche nécessite donc de traiter ces échantillons en deux étapes par l'échange des marqueurs fluorescents (***dye-swap***) afin de corriger ce biais d'incorporation ([Dobbin *et al.*, 2003], [Smyth *et al.*, 2003], [Knapen *et al.*, 2009]). La méthode d'échange de marqueurs fluorescents consiste à inverser le marquage des 2 échantillons et donc à hybrider chaque échantillon 2 fois, après marquage avec chaque fluorochrome, ce qui double ainsi le nombre de puces réalisées. Cela pose un problème de coût et nécessite d'avoir du matériel biologique en quantité suffisante. Cette approche est très souvent utilisée dans le cas des études en cancérologie, car elle permet la comparaison directe sur une puce d'un échantillon pathologique par rapport à un échantillon de référence dit sain. Quant aux puces *one-color*, elles présentent comme principaux avantages la simplicité et la flexibilité de la conception expérimentale : les comparaisons entre les différentes puces d'une expérience sont facilitées, surtout lorsque le nombre d'échantillons est important. De plus cette approche permet de réduire les sources de variabilité lors des tests statistiques grâce à l'utilisation de réplicats biologiques et techniques.

Le consortium MAQC («*MicroArray Quality Control* »), créé en février 2005, a démontré qu'en conditions bien contrôlées, les comparaisons inter- et intra-laboratoires de résultats de puces à ADN indiquent une bonne reproductibilité ([Irizarry *et al.*, 2005], [Shi *et al.*, 2006], [Shi *et al.*, 2010]).

2.1.2 Acquisition des données brutes

Dans le cas des puces AgilentTM, le scanner G2565CA utilise le logiciel de quantification « *Agilent Feature Extraction* » (AFE), afin (1) de quantifier, pour chaque fluorochrome étudié, le signal de fluorescence émis par chaque spot, (2) d'évaluer de façon quantitative la qualité du signal, grâce à la détermination du bruit de fond, (3) de déterminer la qualité du spot par la détection des valeurs extrêmes (outliers) et autres spots saturés. Le logiciel calcule ainsi une valeur normalisée par spot et pour chaque fluorochrome (*gProcessedSignal* et/ou *rProcessedSignal*) puis génère le rapport de qualité et le fichier de résultats en format texte. Il est à noter que pour les puces à ADN *two-colors*, AFE calcule également le ratio (Cy5/Cy3), le logarithme en base 2 de celui-ci pour chaque spot de chaque puce à ADN, ainsi qu'une valeur de probabilité (*p*-valeur). En plus de ce fichier de résultats, AFE génère pour chaque puce un rapport de qualité appelé QCreport. Il est toutefois possible d'utiliser d'autres logiciels ou scanners tels que le GenePix®4400A de Molecular Devices, mais celui-ci retourne les résultats sous un autre format tabulé (gpr).

Une fois les lames scannées, il est nécessaire de passer par une étape de prétraitement et de normalisation des données, à l'aide de programmes dédiés tels que certaines librairies R ou encore le logiciel commercial développé par Agilent « GeneSpring GX ». Ce processus permet de réduire les effets dus aux biais techniques, sans pour autant affecter celui de la variation biologique de l'expression des gènes. Cette étape, qui est indispensable quelle que soit la technologie utilisée, est impossible à automatiser en raison des spécificités inhérentes à chacune de ces technologies.

2.2 Correction des données brutes

2.2.1 Prétraitement des données

Pour les puces à ADN de technologie Agilent™, les logiciel d'analyses, comme AFE, proposent différents critères pour évaluer la qualité du signal de chaque spot. Des biais tels que la variance de l'intensité des pixels du spot, la variance du bruit de fond, la présence de taches ou de spots de taille anormale ou encore un faible rapport signal sur bruit, peuvent être ainsi examinés. Puis l'étape de filtrage des données permet de ne conserver que les spots au dessus d'un seuil de qualité prédéfini pour ne pas fausser les résultats ([Smyth *et al.*, 2003]). **Il est cependant à noter que chaque laboratoire possède sa propre méthode empirique de filtrage des données de puces, car il n'existe pas de méthode standard en ce qui concerne les puces Agilent™.**

C'est à cette étape que le bruit de fond est pris en compte, généralement par soustraction du signal, si celui-ci n'a pas déjà été utilisé par AFE pour générer le signal traité (gProcessedSignal et rProcessedSignal, pour Cy3 et Cy5 respectivement).

2.2.2 Transformation en logarithme base 2

Les données de puce à ADN subissent généralement une transformation logarithmique de base 2 (notée log2) permettant de rendre les ratios (ou en anglais ***fold change***) symétriques et de réduire la dispersion des données en limitant l'influence des valeurs extrêmes. En effet, la plupart des intensités mesurées sont faibles, et donc potentiellement situées au niveau du bruit de fond. Mais surtout, cette transformation permet d'appliquer des tests statistiques paramétriques, car la distribution des valeurs logarithmiques est plus proche de celle d'une loi normale.

2.2.3 Normalisation des données

L'application de méthodes de normalisation est possible dans une certaine mesure : la normalisation suppose que l'effet biologique ne soit pas confondu avec le biais technique que l'on souhaite corriger. Si ce n'est pas le cas, il devient difficile de trancher entre un artéfact technique ou une variabilité biologique. Afin de minimiser au maximum la variabilité expérimentale et pour pouvoir comparer les échantillons entre eux, une normalisation est appliquée aux données dans le but de faire ressortir les différences réellement dues aux variations d'expression des transcrits entre les échantillons.

De nombreuses méthodes de normalisation existent, mais aucune ne peut être appliquée de manière systématique car cela dépend du type de données à étudier. En effet, certaines méthodes de normalisation sont dédiées à une technologie donnée. Par exemple, il existe de nombreuses normalisations dédiées aux puces Affymetrix^{TM}, telles que RMA, MAS 5.0, GCRMA, dChIP, mais celles-ci ne sont pas utilisables pour les données de puces Agilent^{TM}. De plus, les méthodes diffèrent généralement entre des données *one-color* et *two-colors*.

Dans cette partie, seules les principales méthodes de normalisation utilisables pour les puces à ADN de technologie Agilent^{TM}sont décrites.

Il faut déterminer quelle méthode peut corriger au mieux les biais, sans pour autant altérer le signal étudié. Généralement, ces méthodes sont utilisées de la plus simple à la plus sophistiquée si un critère de qualité particulier paraît amélioré.

Le choix de la méthode de normalisation est guidé par des représentations graphiques de type : **nuage de points** (*scatter plot*), **diagramme MA** (*MA plot* ou *Bland-Altman plot*), **histogramme** ou **profil de densité des intensités** ou bien encore **boîte à moustaches** (*box plot*) permettant de visualiser la distribution des données ([Smyth *et al.*, 2003]).

La représentation de type « nuage de points » permet la comparaison de deux échantillons entre eux. On représente ainsi l'intensité de chaque sonde, avec en abscisse le 1er échantillon et en ordonnée le second, les gènes s'éloignant de la diagonale étant différentiellement exprimés.

Le diagramme MA correspond à la visualisation des log ratios en fonction de l'échelle des intensités moyennes. L'abscisse indique l'intensité moyenne des 2 échantillons (A) et l'ordon-

née le logarithme base 2 du ratio des intensités (M) de ces échantillons, et ceci pour chaque sonde avec dans le cas des puces à ADN *two-colors* :

$$\mathbf{M = log\ ratio = log2\ (R/G)\ et\ A = log2((R+G)/2)}$$

où R représente l'intensité de fluorescence rouge (red, celle de la Cy5) et G celle de fluorescence verte (green, la Cy3).

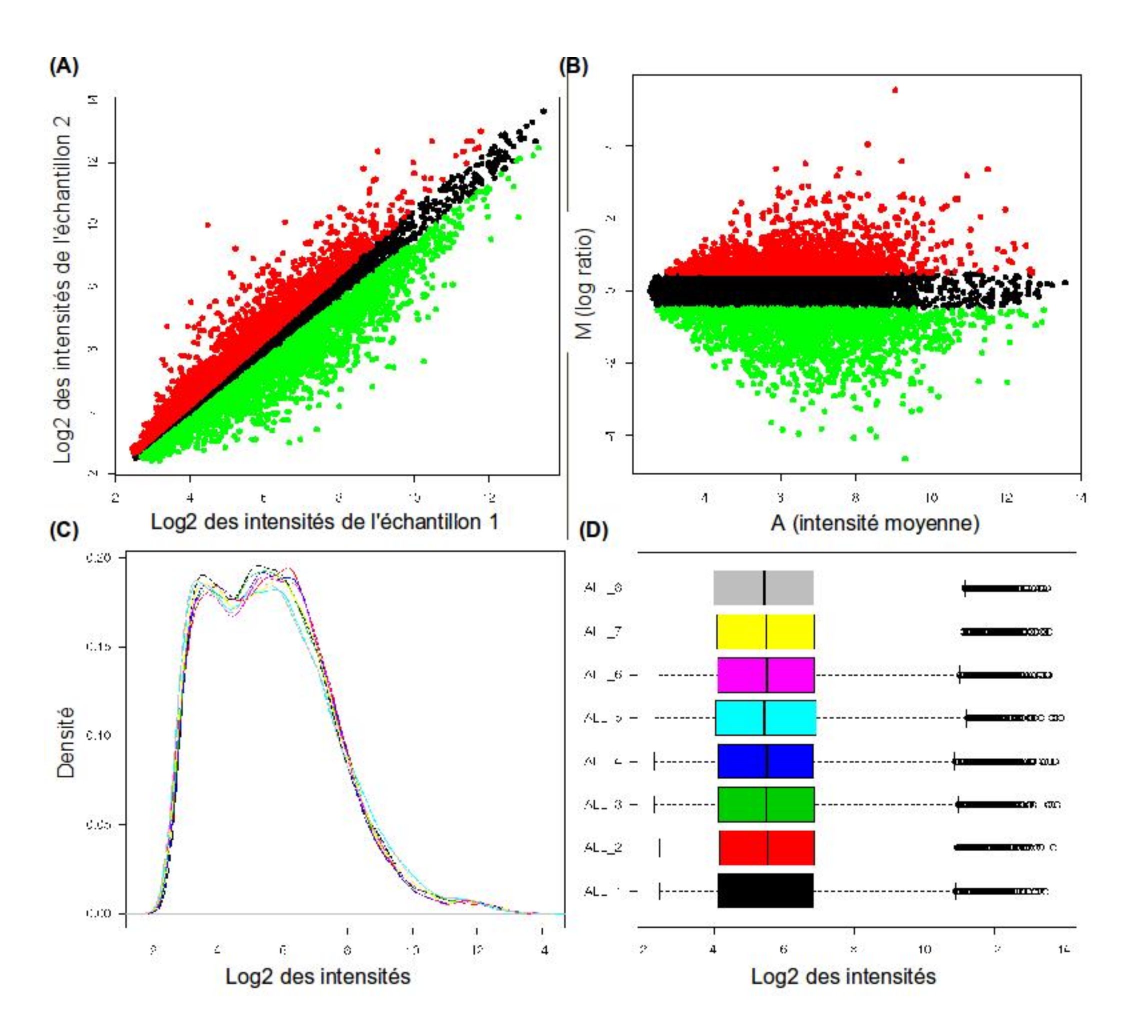


Figure 2.1 – Les différents types de représentations (A) nuage de points, (B) diagramme, (C) histogramme, (D) boîte à moustaches. En rouge sont représentés les gènes sur-exprimés et en vert les gènes sous-exprimés, dans l'échantillon 1 par rapport à l'échantillon 2. Les données utilisées ici sont celles de la librairie R/bioconductor ALL (Acute Lymphoblastic Leukemia, [Chiaretti *et al.*, 2004])

2.2.3.1 Normalisation globale

La méthode de normalisation la plus simple est appelée **normalisation globale par la moyenne ou par la médiane**. Elle consiste à soustraire aux log2 des intensités (ou log ratio) le log2 de la moyenne ou de la médiane des intensités (ou des ratios) de chaque puce. Cette normalisation permet de centrer la distribution des intensités (ou des log ratios) sur 0. Elle reste

néanmoins d'un intérêt limité en raison de la nature souvent non linéaire des relations entre les intensités observées (Ramdas et al., 2001 ; Shoemaker et al., 2001). Ainsi cette méthode ne permet qu'une évaluation de l'erreur systématique pour chaque échantillon, contrôlant des différences proportionnelles à travers les puces.

Il est également possible d'utiliser la **normalisation par centrage et réduction**. Cette technique permet d'uniformiser globalement la distribution des données *two-colors* et *one-color* en centrant les données sur 0 et en fixant l'écart type à 1, puis en calculant les log ratios dans le cas des données *two-colors*. Elle permet ainsi de pouvoir comparer de manière équivalente les différences d'expression de gènes dans plusieurs couples d'échantillons (par exemple tissu tumoral vs. tissu de référence). Le centrage est obtenu en soustrayant aux log ratios la médiane des log ratios de l'échantillon correspondant. Puis les données sont réduites, généralement en divisant les valeurs centrées par l'écart-type de l'échantillon correspondant. Il est également possible d'utiliser la « *Median Absolute Deviation* » (MAD, écart médian absolu), un estimateur plus robuste de la dispersion des données.

Cependant, cette normalisation suppose que les biais observés soient dus à des facteurs globaux concernant tous les gènes (incorporation des marqueurs, qualité de l'hybridation, protocoles expérimentaux...) ; elle ne considère donc aucune région ou effets dépendant des intensités (bruit de fond local, ...), normalement pré-traités précédemment.

2.2.3.2 Normalisation par régression locale

La méthode LOWESS (« *LOcaly WEighted Scatterplot Smoothing* »), proposée par Cleveland en 1979 ([Cleveland, 1979]) et développée par Cleveland et Devlin en 1988, désigne spécifiquement une méthode de régression polynomiale locale pondérée. En fonction du degré du polynôme utilisé, on parle de méthode **LOWESS** ou **LOESS**. Ainsi, pour un degré d'ordre 1 du polynôme, c'est-à-dire une régression linéaire, on parle de méthode LOWESS contrairement au LOESS qui est utilisé pour un ordre 2.

Ce type de normalisation est le plus couramment utilisé pour les puces à ADN *two-colors*. Elle implique que l'expression de la majorité des gènes soit inchangée.

2.2.3.3 Normalisation par les quantiles

La normalisation par les quantiles permet d'uniformiser les distributions des intensités pour un ensemble d'échantillons. Elle est donc particulièrement efficace pour normaliser une série d'échantillons dont les distributions de valeurs d'expression sont proches. Elle suppose que la distribution de l'abondance des gènes soit presque similaire dans tous les échantillons. L'inconvénient de cette méthode est qu'elle peut donner un poids important à des valeurs faibles.

Cette méthode utilise comme référence une puce à ADN dite « synthétique » ([Smyth *et al.*, 2003]). Ces données « synthétiques » correspondent généralement aux

moyennes ou aux médianes des valeurs d'expression calculées sur l'ensemble des échantillons, quantile par quantile. Ces valeurs de référence sont ensuite utilisées pour remplacer, quantile par quantile, les valeurs d'expression (Figure 2.2).

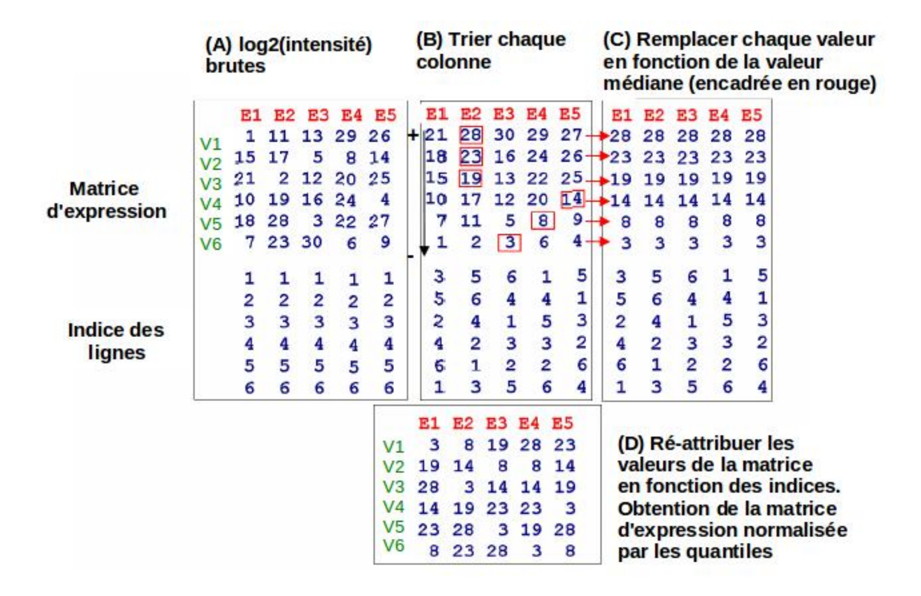


FIGURE 2.2 – Principe de la méthode des quantiles

2.3 Contexte du projet

Constatant l'absence d'outil gratuit disponible permettant d'analyser facilement et rapidement les données obtenues grâce aux puces à ADN de type Agilent^{TM}, nous avons décidé de développer notre propre logiciel.

Ce projet, qui a débuté au cours de mon stage de Master 1 BBSG, porte sur le développement d'une librairie R pour l'analyse de ces données de puce à ADN. Cette librairie, appelée AgiND pour « *Agilent Normalize and Diagnosis* », devait permettre d'extraire les données, mais également de les visualiser de manière très simple et de les normaliser.

L'objectif de cette librairie AgiND n'est pas de proposer une nouvelle méthode de normalisation mais d'offrir les outils permettant d'extraire, de visualiser et de normaliser les données simplement et très rapidement, via l'obtention de fichiers au format texte générés à partir des données brutes extraites à l'aide du logiciel AFE. Cet outil est utilisable tant avec les données de puces *one-color* qu'avec celles des *two-colors*.

Dans la pratique, lorsqu'il s'agit de choisir une méthode de normalisation ou bien d'apprécier la qualité d'un jeu de données, les chercheurs s'aident des représentations graphiques présentées plus haut. Simples et facilement interprétables, ces graphes sont très informatifs et permettent souvent d'orienter les analyses, d'améliorer les protocoles expérimentaux ou même

de définir de nouveaux plans d'expériences.

2.4 Choix du développement d'une librairie R

Le choix du langage R a été motivé par plusieurs aspects. Tout d'abord, **GNU R** (www.r-project.org/) est un langage de programmation, fondé sur le langage S, et un environnement mathématique utilisés pour le traitement de données et l'analyse statistique. Ceci rend cet environnement de travail de plus en plus utilisé par les bioinformaticiens. Il est implémenté dans les langages C, C++, Fortran et Java. R dispose de nombreuses fonctions graphiques et est très régulièrement mis à jour (actuellement 2 nouvelles versions par an).

Cet environnement possèdent plusieurs avantages : (1) c'est un logiciel gratuit et ces sources sont disponibles pour la majorité des systèmes d'exploitation (Windows, Linux et Mac OS), (2) sa syntaxe est intuitive et permet aux biologistes de l'utiliser en ayant recours à des commandes faisant appel à des librairies ou des fonctions préalablement créées ou installées, (3) il permet l'inclusion de fonctionnalités développées dans d'autres langages de programmation tels que le C, le C++, le perl et même des commandes systèmes, (4) il intègre les fonctions dans des paquets ou librairies, accessibles à la communauté scientifique via des dépôts tels que : bioconductor, Comprehensive R Archive Network (CRAN), Omegahat.

Les **librairies R** sont maintenant communément utilisées dans de nombreux domaines scientifiques. Elles sont devenu un outil de travail très apprécié de la communauté bioinformatique, car facile d'accés. Ainsi, de nombreuses librairies dédiées au traitement des données de puces à ADN ont vu le jour. Elles permettent, entre autre :

- l'annotation des données via l'accès à des bases de données ou par la création de librairies contenant l'annotation spécifique à une plateforme de puces à ADN ;
- l'extraction des données à partir de fichiers complexes ;
- la visualisation de ces données via des librairies graphiques ;
- le prétraitement et la normalisation, via différentes méthodes et pour diverses technologies ;
- l'étude statistique, soit à l'aide de tests classiquement utilisés (test t, ANOVA, regression lineaire, SAM), soit grâce à de nouvelles approches et des réanalyses possibles de jeux de données de puces à ADN ;
- le réarrangement et l'export des données.

Ces librairies contiennent des définitions d'objets complexes de type S4 correspondant à une liste de vecteurs et/ou matrices, ainsi que des méthodes et fonctions possédant chacune une fiche d'aide associée. Ces fiches d'aides mais également le manuel d'utilisation de la librairie, appelé vignette sous R, sont écrits en langage LaTeX avec des fichiers « Rd » pour les fiches d'aide et « Rnw » pour la vignette. Ceci permet le balisage et une formulation standard des documents, l'extraction facile de renseignement et l'inclusion de code R au sein de la vignette. Ces librairies ont une structure/architecture définie commune (Figure 2.3).

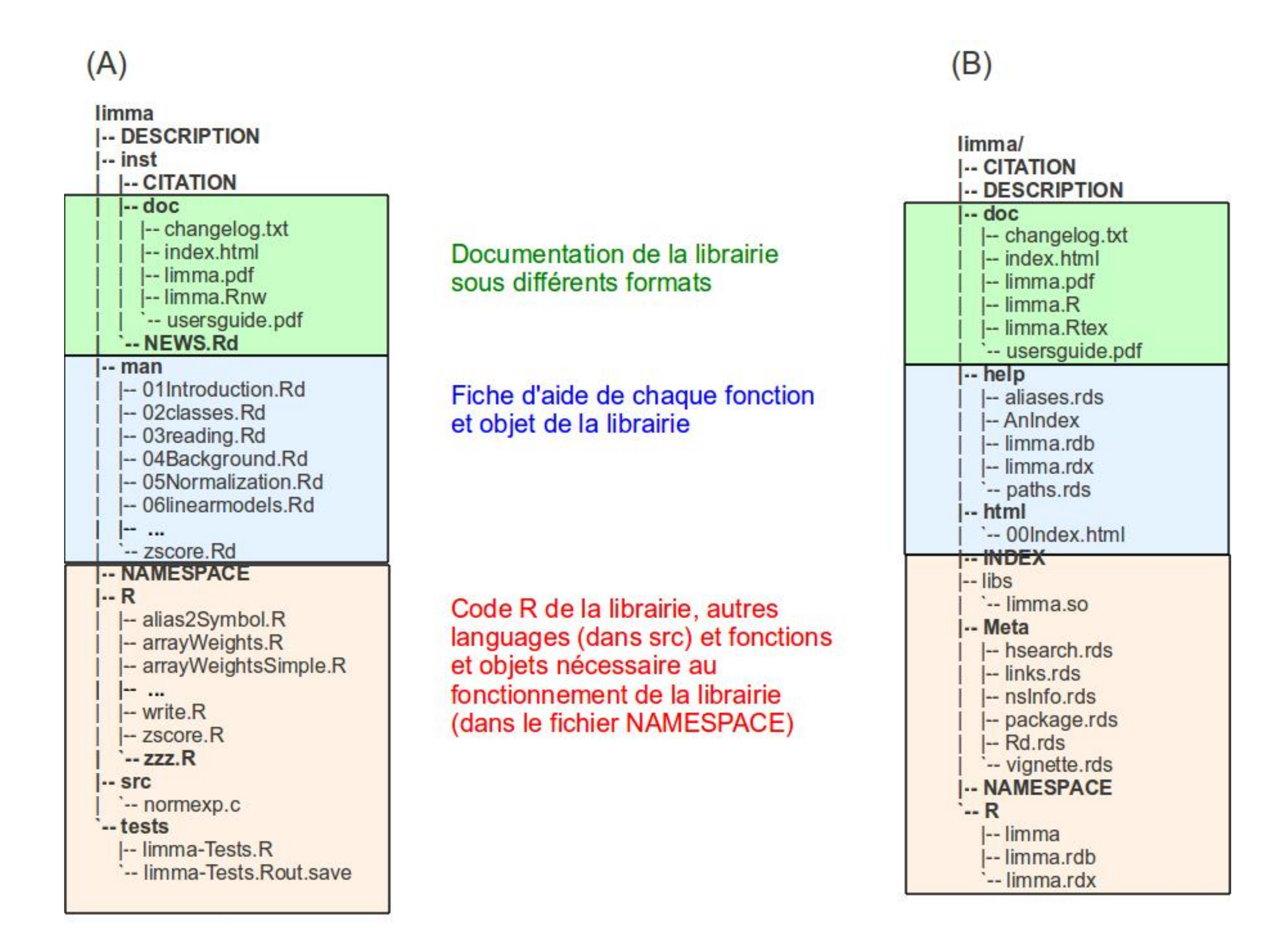


FIGURE 2.3 – Exemple de structure d'une librairie R, ici la librairie R limma avec en (A) son architecture de fichier au niveau du code source, et en (B) celle après compilation et installation de la librairie

2.5 Principe de la librairie R AgiND

Cette librairie est constituée d'un ensemble d'objets et de fonctions codés en R, et d'un programme écrit en langage C permettant l'extraction des données de AFE de manière plus rapide. Ce programme C est appelé par le code R. Enfin, AgiND fait également appel à d'autres librairies R telles que Biobase, limma, marray, geneplotter, annotate, AnnotationDbi et lattice. Il y a également des fiches d'aide pour chaque fonction ou classe d'objet créée ainsi qu'un manuel d'utilisation.

Les données brutes initiales sont obtenues à partir de AFE et correspondent à des fichiers texte tabulés contenant trois tables : deux de paramètres expérimentaux et une de résultats. Ainsi la première table FEPARAMS contient les paramètres d'entrée et les options choisies par l'utilisateur (en accord avec les paramètres du protocole utilisé, comme par exemple « GE1_105_Jan09 »), tandis que la seconde table STATS contient des paramètres déterminés par le scanner. La dernière table FEATURES est un tableau contenant plus de 90 colonnes de résultats avec une ligne par sonde : nom du gène, log ratio, signal normalisé intra-array,

signal moyen, ... Si le fichier correspond à des expériences de puces à ADN *two-colors*, celui-ci contiendra des colonnes pour chacune des couleurs. Dans cette dernière table, seules les informations précédemment citées sont extraites. De plus, un fichier est généré pour chaque échantillon, il est donc nécessaire de combiner ces informations en un seul fichier final pour une expérience donnée.

Pour chaque échantillon, les données, une fois extraites, sont collectées à l'intérieur d'un seul objet de type S4, différent pour les données brutes et normalisées. C'est un objet complexe composé de multiples objets simples : vecteur, scalaire, matrice. Pour les données brutes *one-color*, cet objet est de classe AgilentBatch alors que pour les objets *two-colors*, il se nomme AgilentBatchRG. Une fois les données normalisées, les objets créés sont de classe AgilentNorm et AgilentNormRG, respectivement pour les puces à ADN *one-color* et *two-colors*.

L'étape suivante concerne l'établissement du contrôle qualité des données brutes et normalisées, à l'aide de l'objet qui comporte les principales informations nécessaires (voir le manuel d'utilisation pour une description des données contenues dans l'objet). Dans ce but, des représentations graphiques peuvent être utilisées pour visualiser la distribution :

– des contrôles négatifs : il s'agit d'ADN hétérologue (ou hétéroduplex) correspondant à de l'ADN bicaténaire formé par l'appariement de deux brins d'origine différente ; il présente des domaines en boucle dans les zones où les appariements ne se font pas correctement ;
– des contrôles positifs (« *Spiked-in controls* ») : ce sont des sondes spécifiques d'ADNc de différentes concentrations connues, correspondant à une gamme étalon réalisée par dilutions successives. Ces ADNc sont rajoutés aux échantillons lors du marquage avant l'hybridation sur puce à ADN. Ce sont donc de bons contrôles de qualité et de normalisation des échantillons ;
– des gènes de référence ou gènes de ménage, tels que les protéines ribosomales, dont l'expression est considérée constante dans toutes les conditions, donc dans tous les échantillons.

Cette librairie propose de visualiser les échantillons sous forme :

– de boîtes à moustaches par la fonction agBoxplot ;
– de diagramme MA par la fonction agMAplot ;
– d'images avec la fonction agImage ;
– d'histogramme pour un ou plusieurs gènes d'intérêt par la fonction agPlot.

Des fonctions sont également prévues pour supprimer, selon le besoin de l'utilisateur, certaines des sondes de l'objet de type S4 avant ou après normalisation par la fonction agExclude. Ceci peut également être simplement réalisé en ligne de commande ; en effet, une fonction d'indexation de l'objet de type S4 permet d'extraire toutes les informations de l'objet comme pour une matrice, à partir des indices des colonnes et/ou des lignes à conserver. Il est ainsi possible de créer des sous-groupes d'échantillons et/ou de gènes pour une analyse « personnalisée ».

L'étape de normalisation est cruciale pour pouvoir exploiter les données et repérer d'éventuels biais expérimentaux. Cette librairie propose la transformation en log2 des données et l'accès à deux normalisations : par la méthode des quantiles et par le LOWESS. De plus,

l'utilisateur peut exporter une partie des données au format générique ExpressionSet pour utiliser d'autres librairies R proposant d'autres types de normalisation. Un export des données au format de la librairie R/bioconductor marray a également été implémenté pour permettre l'utilisation des librairies marray et limma.

Enfin les données normalisées peuvent être sauvegardées sous forme de matrice d'expression avec en colonne les échantillons et en ligne les sondes. L'utilisateur peut également choisir à cette étape de supprimer les sondes contrôles.

Par défaut, si aucune information n'est donnée sur les échantillons à l'aide d'un objet de classe AnnotatedDataFrame, les identifiants des échantillons sont directement récupérés à l'aide des noms de fichiers. Seuls le code barre et la position sur la puce à ADN seront conservés ultérieurement. Le fichier « US83700202_252800413012_S01_GE1_107_Sep09_1_1.txt » portera par exemple le nom simplifié « 252800413012_1_1 ».

Pour les gènes, les Feature number et le symbole des gènes sont séparés par des « | », par exemple « 4||U2AF1L4 ». Ceci permet l'obtention d'un identifiant unique puisqu'un transcrit peut être présent à plusieurs reprises sur la puce à ADN. Il est également possible d'y rajouter par programmation les références des sondes, les descriptions de gènes et d'autres informations.

Un résumé de l'installation de la librairie AgiND et de son utilisation est présenté dans le manuel d'utilisation téléchargeable avec la librairie (voir Annexe A).

Non publiée à l'heure actuelle, cette librairie est déjà utilisée en routine par les utilisateurs de la plateforme transcriptome TGML. Par ailleurs, elle a permis plusieurs collaborations qui sont présentées dans le chapitre suivant de ce manuscrit. Cette librairie n'a pas été soumise à des dépôts de librairies R, comme bioconductor, cependant elle est mise à disposition sur le site web du laboratoire : http://tagc.univ-mrs.fr/AgiND

2.6 Discussions et Perspectives

Au début de ce projet, les seuls outils gratuits disponibles étaient les librairies R marray et limma ([Smyth, 2005]). Elles permettent de contrôler la qualité des puces à ADN de technologie Agilent^{TM}et de normaliser les données générées. Mais elles présentaient deux inconvénients : elles chargeaient très lentement les données et n'étaient pas spécifiques de l'analyse des puces à ADN de type Agilent^{TM}, ce qui a nécessité le développement de notre propre librairie R AgiND.

Pour y remédier, les bioinformaticiens ont du développer des librairies R dédiées permettant soit d'estimer la qualité des puces à ADN d'une expérience, soit de normaliser ces données (Table 2.1). Ces librairies R ont été mises à disposition sur le site du CRAN ou de bioconductor : arrayQuality ([Paquet & Yang, 2008], Agi4x44PreProcess ([Lopez-Romero, 2008]), arrayQualityMetrics ([Kauffmann & Huber, 2008]), agilp ([Chain *et al.*, 2010]), BABAR ([Alston *et al.*, 2010]), GOULPHAR (pipeline de l'ENS, Paris, [Lemoine *et al.*, 2006]) ou bien encore des plugins pour Excel comme arrayTools. En parallèle, les librairies de base de

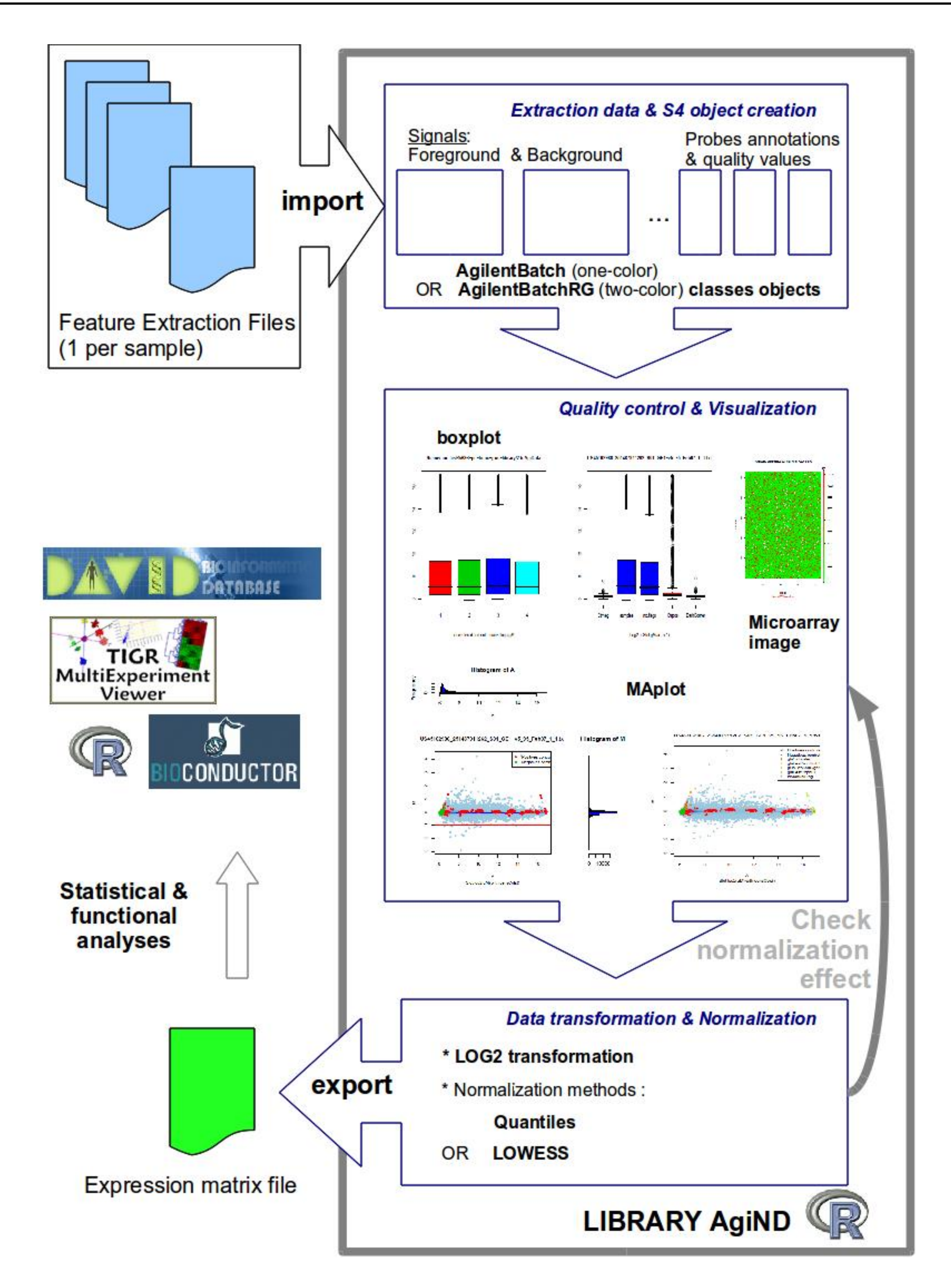


FIGURE 2.4 – Schéma récapitulatif de l'analyse de données de puces à ADN incluant la librairie AgiND

bioconductor pour l'analyse des puces à ADN, marray et limma, ont été améliorées.

Les puces à ADN de type Agilent^{TM}sont utilisées dans près de 25% des expériences réalisées à partir d'échantillons humains soumises à GEO. L'analyse des données générées

représente un véritable enjeu, obligeant les entreprises commerciales, à commencer par Agilent^{TM}, à créer des logiciels dédiés (GeneSpring GX).

La librairie que j'ai mise au point est utilisable avec tous les puces compatibles avec le format AFE, et n'est donc pas uniquement dédiée aux puces 4x44k comme l'est Agi4x44PreProcess. En effet, AgiND utilise les paramètres présents au début du fichier pour générer et vérifier la taille des objets. Elle peut ainsi être utilisée tant pour les puces 4x44k, que pour les nouvelles puces 8x60k par exemple.

De plus, AgiND est utilisable aussi bien pour les données de puces *one-color* que pour les *two-colors*. Enfin, la génération d'un format de sortie de classe ExpressionnSet permet l'utilisation d'autres librairies R comme vsn, afin de normaliser les données brutes extraites par AgiND, ou bien encore d'utiliser des librairies graphiques telles que arrayQualityMetrics.

Le développement d'AgiND n'est pas terminé. On peut encore améliorer cette librairie. Pour permettre son utilisation avec tous les types de formats de fichiers bruts, il faudrait permettre l'importation de données obtenues à l'aide du logiciel GenePix, comme c'est le cas des librairies BABAR, GOULPHAR ou bien encore limma. De plus, il faudrait inclure des fonctions vérifiant la qualité des données de manière plus ou moins automatique, comme l'étude des contrôles positifs et négatifs, et générant des graphiques. Enfin, un rapport d'analyse comme sous ArrayQualityMetrics pourrait être généré et donné aux clients de la plateforme TGML. Ceci serait un gage supplémentaire de qualité en vue d'une analyse de données.

Logiciel ou librairie R	Type de fichier d'entrée		Contrôle qualité	Filtrage des données	Normalisation(s)	Langage(s)
	AFE	GenePix				
AgiND	**X**		**X**	**X**	**Quantiles, lowess**	**R, C**
limma	X	X	X	X	Quantiles, lowess, print-tip loess, scale	R
Agi4x44PreProcess	X		X	X	Quantiles, vsn	R, C++,C, fortran
GOULPHAR		X	X	X	Lowess	R
agilp	X		X	X	Lowess	R
BABAR		X	X	X	Lowess	R
arrayQuality			X			R
arrayQualityMetrics			X			R
genefilter				X		R
vsn					vsn	R
arrayTools	X		X	X		Excel, R, Java
GeneSpring GX	X	X			Quantiles, lowess, print-tip loess, scale	Applet JAVA, API SOAP, API JYTHON (JAVA Based Python), R

TABLE 2.1 – Synthèse des principaux outils permettant l'analyse des données de puces à ADN de technologie Agilent™. En grisé, le logiciel commercial développé par Agilent et en gras, les caractéristiques de notre librairie R, AgiND

Chapitre 3

Analyses de données de puces à ADN

Sommaire

Une fois les données normalisées, l'objectif est de mettre en évidence des gènes différentiellement exprimés dans les échantillons. Ceci peut être obtenu par le simple calcul des amplitudes de variations d'expression (ratio), mais c'est généralement insuffisant. Des approches statistiques sont donc nécessaires afin d'estimer et de distinguer la variabilité intra- et inter-groupe. De nombreux tests statistiques ont ainsi été proposés, allant du test t de Welch aux approches bayésiennes en passant par les analyses de variance. Ces méthodes ont été utilisées lors de collaborations en vue d'obtenir des signatures moléculaires dans le cas d'études de pathologies.

3.1 Sélection de gènes

L'application des tests dépend de plusieurs paramètres, mais ceux-ci sont tous appliqués à chaque gène présent sur la puce à ADN afin de déterminer les gènes différentiellement exprimés en fonction des différents groupes d'échantillons. En effet, il faut distinguer les cas où les données analysées sont indépendantes ([Golub *et al.*, 1999]) ou appariées ([Perou *et al.*, 2000]). Il est également important d'évaluer la distribution des données, pour déterminer si l'on peut utiliser des tests paramétriques ou non paramétriques. On parle de tests paramétriques (par exemple : t-test, ANOVA) lorsque l'on postule que les données sont issues d'une distribution paramétrée (distribution normale par exemple). L'hypothèse de normalité sous-jacente des données est souvent utilisée ; en effet la transformation des données en logarithme de base 2 permet d'obtenir une distribution assimilable à une gaussienne. La moyenne et la variance de ces données suffisent ainsi à caractériser complètement leur distribution. Contrairement aux tests paramétriques, en ne faisant aucune hypothèse sur la distribution des données, les tests non-paramétriques élargissent le champ d'application des méthodes statistiques. En contrepartie, ils sont moins puissants lorsque ces hypothèses sont compatibles avec les données.

Les résultats des tests statistiques multiples doivent être corrigés pour minimiser le nombre de faux positifs. L'hypothèse nulle, notée H0, de ces tests est qu'il n'existe pas de différence significative (pas d'effet biologique) entre 2 ou plusieurs groupes au risque α de se tromper. Le résultat du test d'hypothèse est une probabilité, désignée p-valeur, qui croît plus le phénomène observé (la variation de l'expression d'un gène entre 2 ou plusieurs conditions) est lié au hasard. Le principal mode de contrôle de l'erreur de type I (ou risque α) est le FDR (« False Discovery Rate »). Il permet d'estimer la proportion q d'erreurs parmi les gènes considérés comme différentiellement exprimés (faux positifs). Les méthodes FDR sont généralement plus puissantes et moins conservatrices que les autres approches telles que le « Family wise error rate », FWER. Selon le seuil choisi par l'utilisateur, la sélection de gènes différentiellement exprimés sera plus ou moins sévère ; généralement, un seuil standard de FDR à 5% est utilisé.

3.1.1 Test t

Le test t de Student compare les moyennes de deux groupes d'échantillons et détermine, en fonction d'un risque fixé, si ces moyennes sont significativement différentes pour chaque gène ([Callow *et al.*, 2000]). Ce test paramétrique peut être réalisé de manière appariée ou non. Les tests appariés sont plus puissants car en couplant par échantillon (dans le cas où un même échantillon est utilisé avant/après un traitement, par exemple), cela permet de réduire la variabilité de l'expression des gènes qui diffère d'un échantillon à l'autre. Ceci est généralement utilisé pour réduire ou éliminer les effets de facteurs confondants indépendants de l'appartenance à un des deux groupes. Un test t non apparié peut être normalement utilisé lorsque deux ensembles distincts d'échantillons indépendants et identiquement distribués sont comparés. Un des pré-requis de ce test est que la variance des deux échantillons est identique (homoscédasticité).

Contrairement au test t de Student, le test t de Welch tient compte de l'inégalité de la

variance des deux groupes d'échantillons et peut donc être mis en oeuvre lorsque l'hypothèse d'homoscédasticité n'est pas vérifiée, ce qui est souvent le cas pour les données de puces à ADN.

Le test t de Welch définit la valeur statistique t par la formule suivante :

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

où $\bar{x}$, s et n correspondent, dans le cas des puces à ADN, respectivement à la moyenne des intensités d'un gène, à l'écart-type et à la taille des groupes d'échantillons A ou B.

Pour chaque gène, une p-valeur est estimée soit à partir de la distribution de la statistique t (c'est-à -dire de la loi de Student) (Figure 3.1), soit à partir de permutations, ce qui permet de définir un FDR.

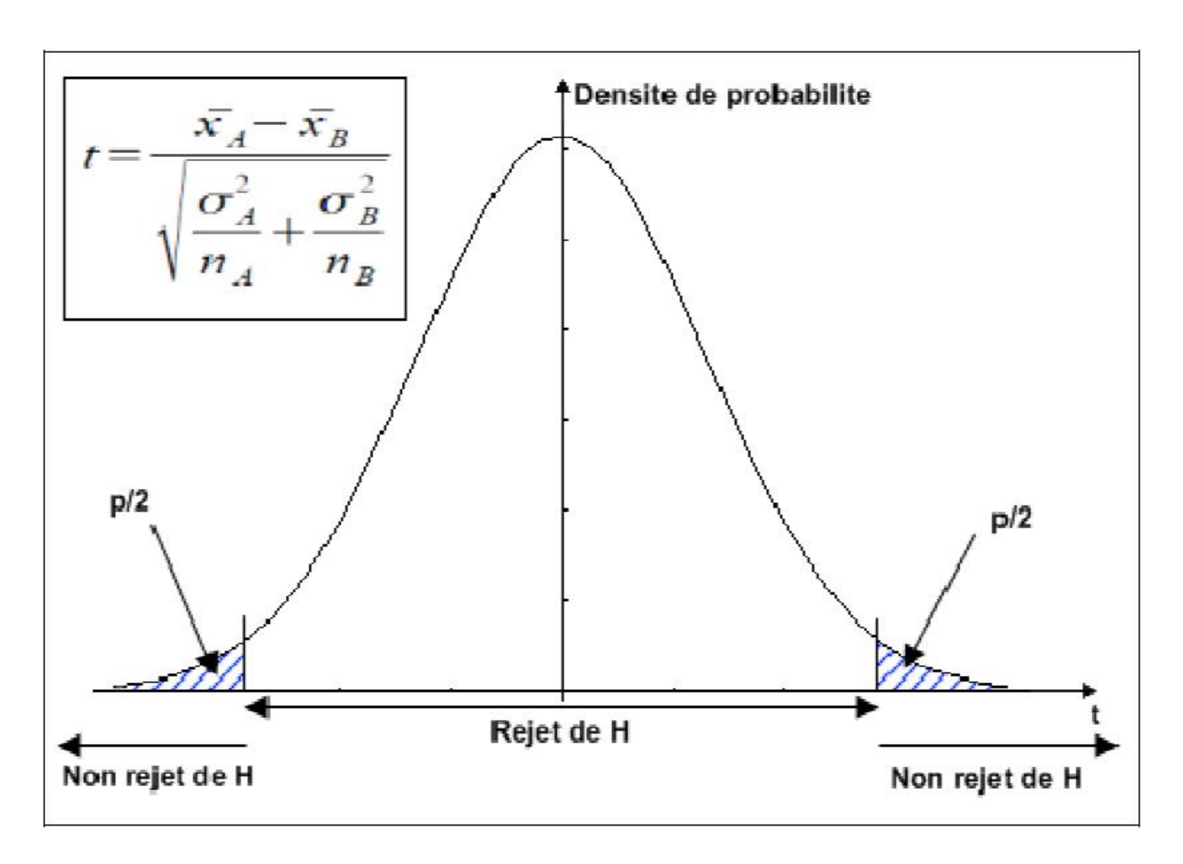


FIGURE 3.1 – Distribution de la loi de Student

Une représentation graphique simple permet d'interpréter les résultats du test t, le volcano plot (Figure 3.2). Ce graphique représente en ordonnée le logarithme de base 10 des p-valeurs provenant du test t et en abscisse le logarithme de base 2 du fold change. Les gènes différentiellement exprimés sont ceux présentant de faibles valeurs p (soit des valeurs en ordonnée les plus élevées possibles, car elles sont en log10) et affichant des valeurs absolues de fold change élevées. Généralement, un seuil arbitraire est appliqué pour conserver les valeurs supérieures à 1 (soit une expression deux fois plus importante dans un des échantillons).

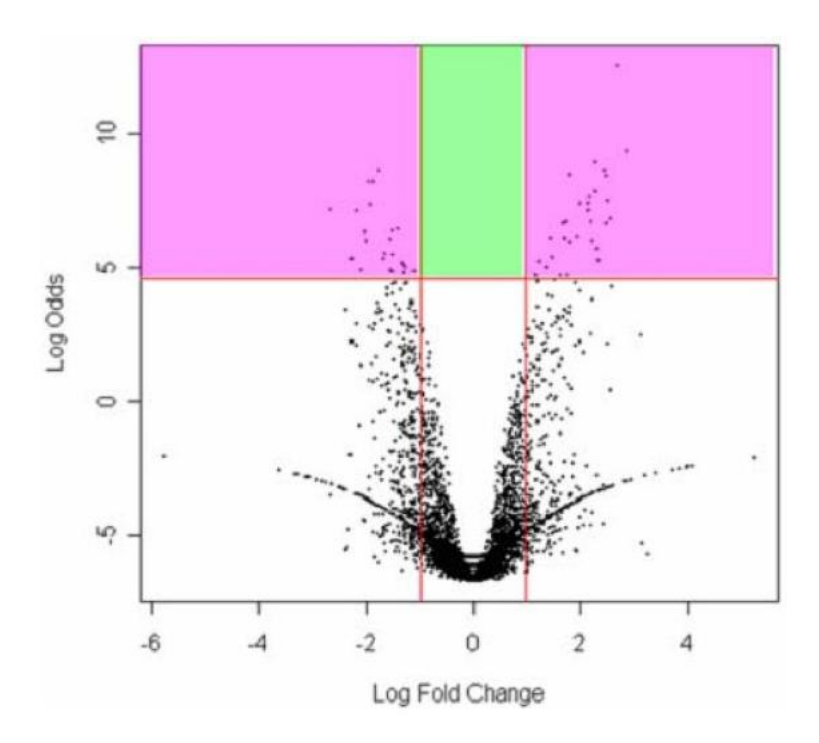


FIGURE 3.2 – Représentation de type volcano plot. Chaque point correspond à un gène. Deux filtres sont réalisés un sur chaque axe, pour déterminer les gènes discriminant. Les zones en rose correspondent à des régions dans lesquelles les gènes ont un log ratio supérieure ou égal à 1 et dont la *p*-valeur du test t est inférieur à 10^{-5}

3.1.2 Significant Analysis of Microarrays (SAM)

La méthode SAM est un test non-paramétrique, qui permet d'identifier des gènes différentiellement exprimés entre deux groupes d'échantillons, et ceci sans a priori sur leur distribution. SAM assigne un score à chaque gène, sur la base du changement d'expression relatif des gènes par rapport à la déviation standard des réplicats. Pour cela, il utilise des permutations dans les mesures pour estimer le FDR. La statistique du test SAM d est définie par ([Tusher *et al.*, 2001]) :

$d(i) = \frac{\bar{x}_A(i)-\bar{x}_B(i)}{s(i)+S_0}$

où xA et xB représentent les moyennes d'expression pour le gène i dans les conditions A et B, et s(i) est une estimation de la variance représentant l'écart-type pour le gène i tel que :

$s(i) = \sqrt{a(\sum m[x_m(i) - \bar{x}_A(i)]^2 + \sum n[x_n(i) - \bar{x}_B(i)]^2)}$

et

$a = \frac{\frac{1}{n_A}+\frac{1}{n_B}}{n_A+n_B-2}$

avec xm et xn les sommes des expressions pour les échantillons du groupe A et B respectivement, et nA et nB le nombre d'échantillons dans les groupes A et B.

Cette statistique est sensiblement identique à celle du test t. La différence essentielle est la présence au dénominateur d'un facteur correctif S_0, Ce facteur correspond à une faible valeur positive calculée pour minimiser l'effet de la variance. En effet, les gènes peu abondant ont une variance faible. Ce score est calculé pour chaque gène en fonction des deux groupes fournis

par l'utilisateur. Ce même calcul est réalisé à partir des données initiales par un nombre défini de permutations d'un ensemble d'échantillons, afin de générer une distribution simulée des valeurs d_E. Les valeurs de d observées sont ensuite comparées aux valeurs simulées d_E (Figure 3.3). Les gènes différentiellement exprimés sont alors sélectionnés en fonction du FDR calculé à partir des permutations. Un seuil arbitraire de FDR est choisi par l'utilisateur (généralement 5%) en fonction du nombre de faux positifs qu'il est prêt à accepter.

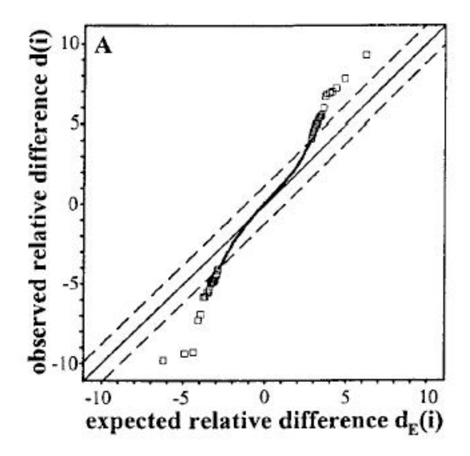


FIGURE 3.3 – Représentation de la valeur de d obtenue pour chaque gène i, soit d(i), en fonction de la valeur simulée d_E(i).

3.1.3 ANalysis Of VAriance (ANOVA)

L'analyse de variance (ANalysis Of Variance, ANOVA) est une méthode paramétrique utilisée pour comparer les moyennes d'au moins 3 groupes d'échantillons ([Draghici *et al.*, 2003]). Elle postule que les données sont distribuées suivant une loi normale et que les gènes sont indépendants, pour une analyse globale. Ceci n'est généralement pas le cas pour les données de puces à ADN, car les gènes ne sont pas indépendants en terme de régulation.

Néanmoins, elle est couramment utilisée pour estimer, si les différences observées entre ces moyennes sont significatives ou non, en prenant en compte diverses variables explicatives catégorielles (type cellulaire, temps de traitement, sexe). Selon le nombre de facteurs pris en compte, on parle d'ANOVA mono ou multi-factorielle. Son principe est basé sur le fait que la variance observée est due à la contribution d'une ou de plusieurs sources distinctes.

Considérons par exemple le cas de cellules provenant de différents types de cancers (C={'sein','prostate','ovaire'}), de patients de sexes différents (S={'Femme','Homme'}) qui peut être représenté par la formule :

$$log(Y_{ijkl}) = \mu + G_i + E_j + T_k + S_l + (GE)_{ij} + (GT)_{ik} + (GS)_{il} + (ET)_{jk} + (ES)_{jl} + (TS)_{kl} + \xi_{ijkl}$$

où G, E, T, S représentent respectivement les effets dus aux gènes, aux échantillons, au type de cancer, au sexe. Les interactions entre deux de ces facteurs sont notées entre parenthèses

comme par exemple pour $(TS)_{kl}$ qui correspond donc à l'interaction entre le type de cancer k et le sexe s. Les gènes différentiellement exprimés seront ceux pour lesquels l'interaction avec les échantillons (GE) auront les plus faibles p-valeurs.

3.2 Méthodes de classification non supervisées

La classification de gènes ou d'échantillons peut être obtenue par (1) des méthodes supervisées, si l'on tient compte de l'expression différentielle des gènes dans différents groupes d'échantillons selon leur phénotype, ou (2) non supervisées c'est-à dire sans a priori en se basant sur l'ensemble des échantillons.

Diverses méthodes de classification non supervisées ont été appliquées à l'identification des profils dans les données d'expression géniques. Elles peuvent être classées en 2 catégories : les méthodes de regroupement (classification hiérarchique) et les méthodes de partitionnement (k-moyens, cartes auto-adaptatives) en n groupes de gènes ou clusters.

Divers outils gratuits permettent de générer, ou uniquement de représenter ces classifications : TMeV, cluster, Treeview, fonctions et librairies R.

3.2.1 La méthode de classification hiérarchique

La classification hiérarchique a pour avantage d'être simple à mettre en oeuvre et son résultat peut être facilement visualisé. Elle est devenue l'une des méthodes les plus utilisées pour l'analyse des données d'expression génique. C'est une approche d'agglomération dans laquelle les profils d'expression sont simplement regroupés sur la base de leur similarité. Les groupes ainsi obtenus sont ensuite joints jusqu'à ce que le processus ait été mené à son terme, en formant un seul arbre hiérarchique appelé aussi dendogramme (Figure 3.4).

La classification hiérarchique permet de représenter la matrice des intensités d'expression normalisées afin de visualiser simplement les gènes ayant des profils similaires. Cette matrice, dont la distribution est médiane-centrée sur les gènes, est telle que chaque colonne correspond à une expérience et chaque ligne correspond à la sonde d'un transcrit. On représente généralement les ratios ou valeurs d'intensités normalisées grâce à une échelle de couleurs allant du vert (gènes réprimés) au rouge (gènes induits). Cette représentation est communément appelée heatmap (Figure 3.4).

3.2.2 La méthode des k-moyens (k-means)

Dans la méthode de partitionnement k-means, les éléments sont répartis en un nombre fixe k de clusters (où k est spécifié par l'utilisateur). Dans un premier temps, tous les gènes sont assignés au hasard à l'un des k clusters ; puis, un vecteur d'expression moyen est calculé pour

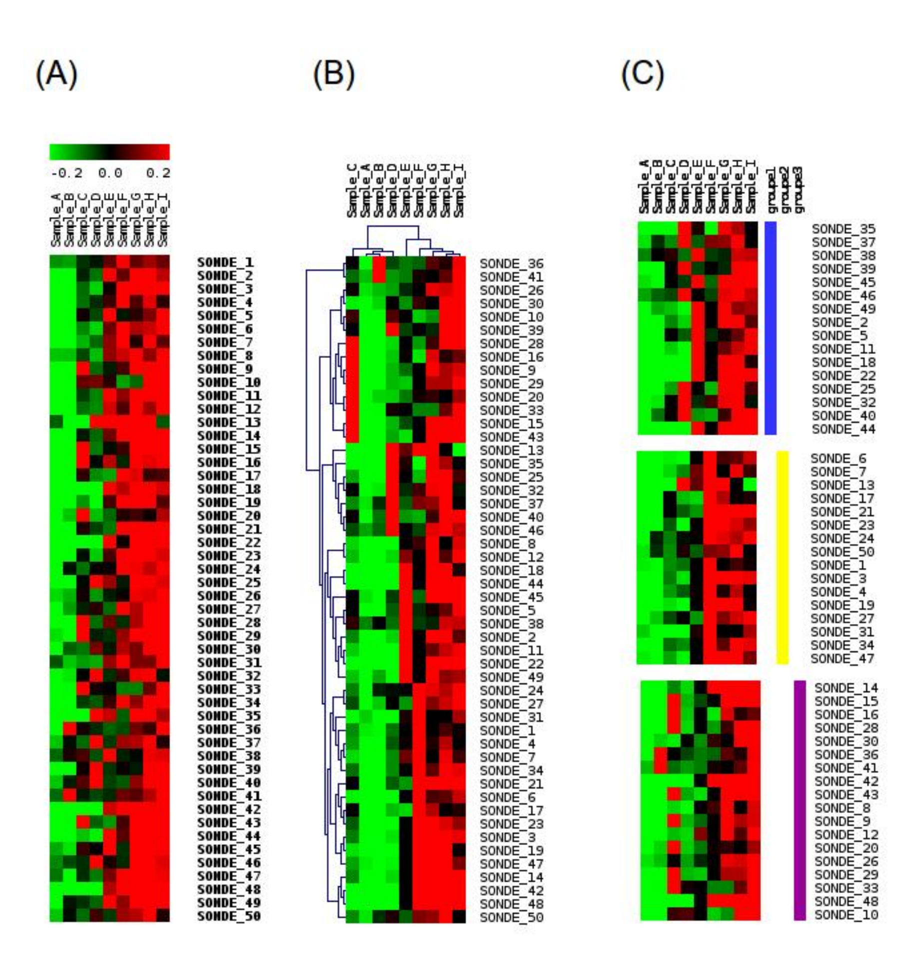


FIGURE 3.4 – Représentation d'un jeu de données : (A) le heatmap sur les données brutes, (B) la classification hiérarchique, et (C) la partitionnement par la méthode k-means (k=3)

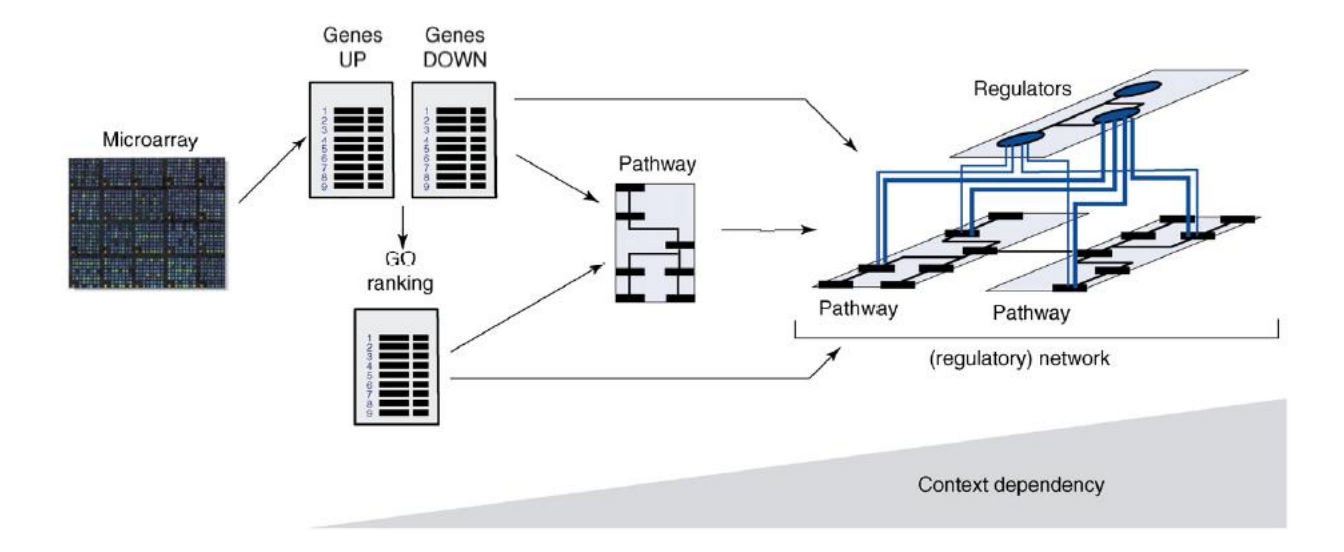


FIGURE 3.5 – Interprétation des gènes différentiellement exprimés à partir d'une expérience de puces à ADN. Cette interprétation dépend de l'étude menée et permet la génération de réseaux de gènes contextualisés ([Werner, 2008])

chaque groupe. Il permet de déterminer les distances entre clusters. De manière itérative, les éléments sont déplacés d'un cluster à l'autre et les distances sont recalculées à chaque itération. Les éléments ne sont autorisés à rester dans le nouveau cluster que si celui-ci est plus proche de l'élément que le cluster précédent, ceci afin de minimiser la somme des carrés à l'intérieur de chaque cluster.

3.2.3 Self-organizing maps (SOM)

Cette méthode appelée en français « Carte auto-adaptative » est un réseau de neurones artificiels fondé sur des méthodes d'apprentissage non-supervisées. On la désigne souvent par le terme anglais « *Self Organizing Map* » (SOM), on encore carte de Teuvo Kohonen du nom du statisticien ayant développé le concept en 1984 ([Kohonen, 1997], [Tamayo, 1999]). Elle est utilisée pour classifier des données dans un espace multi-dimensionnel, comme dans le cas des puces à ADN.

3.3 Annotation fonctionnelle

Après avoir identifié des groupes de gènes différentiellement exprimés et afin de pouvoir interpréter les données, il est nécessaire de procéder à des tests d'enrichissement fonctionnel. En effet, les gènes co-exprimés sont généralement impliqués dans des processus ou voies de signalisation similaires ([Eisen *et al.*, 1998]) (Figure 3.5).

3.3.1 Les différentes sources d'information

Il existe diverses sources d'information utiles pour l'annotation et donc pour l'interprétation des données de puces à ADN. En effet, de très nombreuses bases de données stockent des informations sur la fonction, la localisation, l'expression tissulaire, la régulation et les interactions des gènes ou de leurs produits (Tableau 3.1). En effet, on considère ici que les transcrits identifiés précédemment sont traduits de manière équivalente en quantité de protéines fonctionnelles. Cela ne tient donc pas compte des mécanismes de régulation post-transcriptionnelle et post-traductionnelle.

Parfois, les données sont organisées en un ensemble structuré de termes et concepts au vocabulaire contrôlé, appelé ontologie. Les concepts sont organisés dans un graphe dont les relations peuvent être des relations sémantiques ou des relations d'inclusion. L'objectif premier d'une ontologie est de modéliser un ensemble de connaissances dans un domaine donné.

L'ontologie la plus connue pour l'annotation de données provenant de puces à ADN est Gene Ontology (GO ; [Ashburner *et al.*, 2000]). Celle-ci propose un vocabulaire contrôlé de termes décrivant les propriétés des produits des gènes. Elle est composée de 3 domaines :

– compartiment cellulaire, ou *cellular component*, décrivant la localisation des protéines au sein de la cellule (comme par exemple : noyau, cytoplasme, membrane) ;
– fonction moléculaire ou *molecular function*, décrivant les activités au niveau moléculaire, telles que la liaison (par exemple le terme GO « *transcription factor binding* », GO :0008134) ou la catalyse ;
– processus biologique ou *biological process*, représentant l'ontologie la plus intéressante pour connaître la fonction des protéines. Elle nous renseigne sur les processus dans lesquels des protéines sont impliquées, comme par exemple la transcription (terme « *transcription, DNA-dependent* », GO :0006351, Figure 3.6).

3.3.2 Quelques outils d'annotation

Plusieurs outils utilisant cette ontologie ont été créés comme AmiGO, GOToolsBox ([Martin *et al.*, 2004]), FATIGO ([Al-Shahrour *et al.*, 2007]). Les autres bases de données ont également mis en place un système permettant des recherches en fonction d'un gène, d'une protéine, d'un processus biologique ou d'une voie de signalisation. D'autres approches sont également utilisées pour obtenir des informations sur des gènes telles que des outils de fouille de texte comme Chilibot ([Chen & Sharp, 2004]), iHOP ([Good *et al.*, 2006]). Enfin, des logiciels proposent également l'accès à différentes sources de données précédemment citées. Parmi les outils gratuits, ceux principalement utilisés par les biologistes et les bioinformaticiens sont « *The Database for Annotation, Visualization and Integrated Discovery* » DAVID knowledgebase ([Huang *et al.*, 2009]) et « *Gene Set Enrichment Analysis* » GSEA ([Subramanian *et al.*, 2005])(Figure 3.7). La base de données DAVID propose ainsi un outil de regroupement d'annotations fonctionnelles permettant l'identification de groupes d'annotations significativement surreprésentées dans une sélection de gènes (Huang da et al. 2007 ; Sherman et al. 2007). Alors que GSEA est une méthode non paramétrique qui détermine si un jeu de gènes défini a priori possède des différences statistiquement significatives entre

Disease	GENETIC ASSOCIATION OMIM DISEASE
Gene Ontology (GO)	GO Biological Process (BP) GO Cellular Component (CC) GO Molecular Function (MF)
General annotations	CHROMOSOME CYTOBAND
Litterature	GENERIF SUMMARY PUBMED ID
Pathways	BIOCARTA KEGG PANTHER REACTOME
Protein domains	BLOCKS COG INTERPRO PFAM SCOP SMART SSF TIGRFAMS
Protein interactions	BIND NCICB CAPATHWAY REACTOME TFBS conserved
Tissues espressions	CGAP EST QUARTILE CGAP SAGE QUARTILE GNF U133A QUARTILE PIR TISSUE SPECIFICITY UNIGENE EST QUARTILE UP TISSUE

Table 3.1 – Liste des principales annotations contenues dans l'outil DAVID knowledgebase, regroupées par domaine

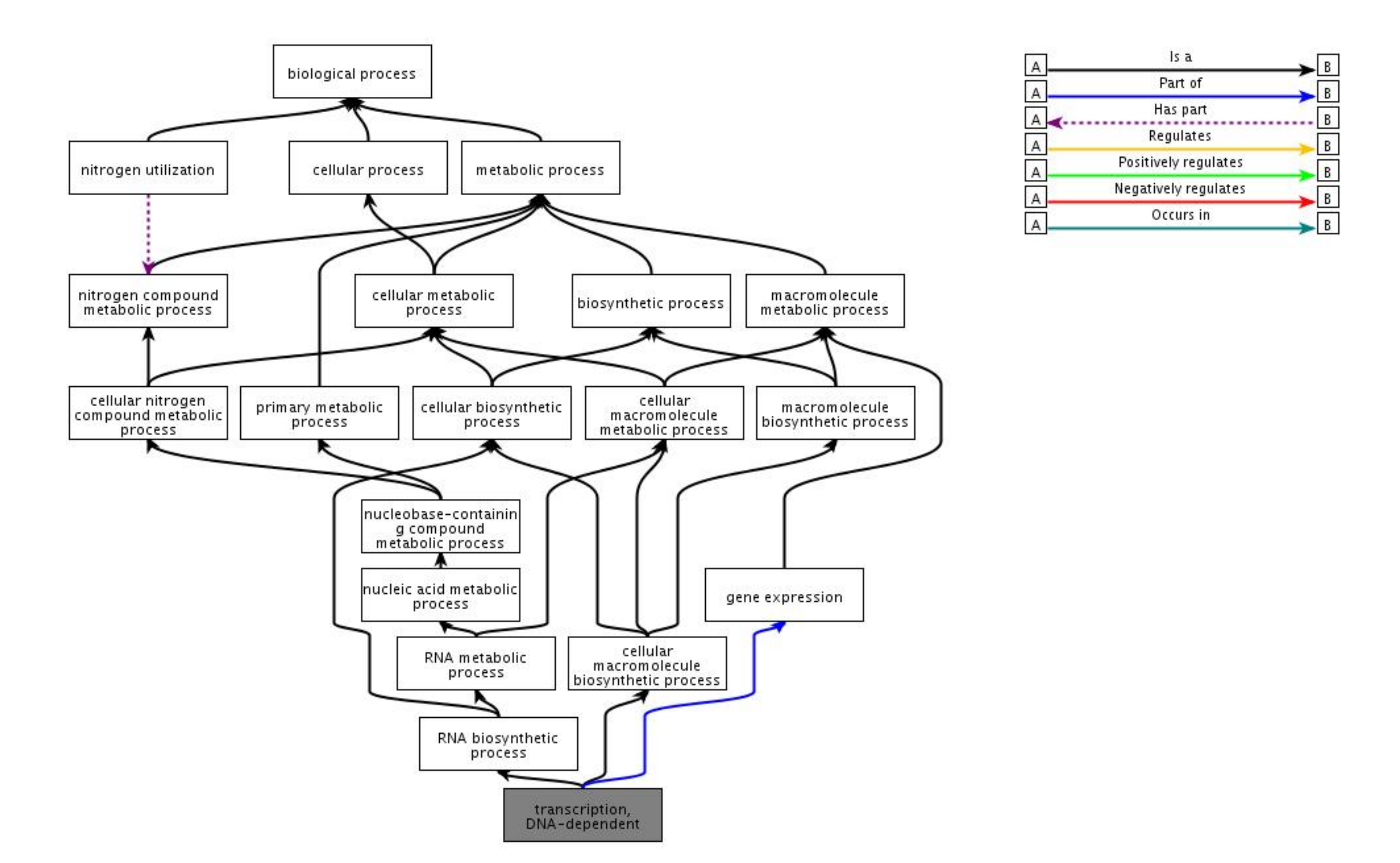


FIGURE 3.6 – Exemple de structure de l'ontologie Gene Ontology « *Biological Process* ». Cette figure schématise les termes parents du terme « *transcription, DNA-dependent* », obtenue à l'aide de l'outil QuickGO (http://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0006351)

deux états biologiques ; cette méthode permet de calculer des scores d'enrichissement fonctionnel en utilisant la base de données moléculaire *Molecular Signature DataBase* (MSigDB) ([Subramanian *et al.*, 2005]). Une application commerciale existe également : *Ingenuity Pathway Analysis* (IPA) qui comporte des annotations vérifiées par des scientifiques et permet la visualisation des gènes surexprimés (en rouge) et sousexprimés (en vert) sous forme de réseaux de gènes contextualisés (Figure 3.7).

3.3.3 Tests d'enrichissement fonctionnel

Lors de l'étape d'annotation fonctionnelle, il ne suffit pas seulement de savoir quelle voie de signalisation ou quelle annotation caractérise au moins l'un des gènes présents dans la liste de gènes différentiellement exprimés, il faut également savoir si l'association d'une partie des gènes de cette liste à une annotation donnée est significative ([Draghici *et al.*, 2003]). Un test d'enrichissement fonctionnel permet de comparer la liste de gènes différentiellement exprimés aux gènes impliqués dans une voie de signalisation ou associés à une annotation fonctionnelle particulière, pour vérifier si la liste de gènes testée est statistiquement enrichie en gènes associés à ce processus biologique ou voie de signalisation.

Plusieurs méthodes statistiques peuvent être utilisées pour cette comparaison ([Draghici *et al.*, 2003]). Elles peuvent inclure (1) le test du chi 2, (2) le test exact de Fisher, (3) la distribution hypergéométrique et (4) le test binomial.

Le test du chi 2 est simple à calculer, mais il ne donne qu'une valeur approximative p, et il est limité aux cas où le nombre d'observations de chaque type (par exemple, les gènes surexprimés qui apparaissent dans le mot clé) est supérieur à cinq. S'il y a moins de cinq observations, une alternative à ce test est celui du test exact de Fisher. Ce calcul correspond à la probabilité exacte de voir le nombre observé d'occurrences. Sinon, la probabilité de l'existence d'un nombre spécifique de gènes d'une classe dans une liste de gènes peut être calculée grâce à la distribution hypergéométrique. Cette distribution est utilisée pour l'échantillonnage de populations finies, mais se rapproche de la distribution binomiale pour un nombre élevé d'échantillons. Étant donné que les puces contiennent généralement des sondes représentant des dizaines de milliers d'ARNm, cette approximation binomiale peut être utilisée. Ces tests statistiques donnent des p-valeurs qui décrivent la probabilité d'obtenir le résultat observé.

Des permutations et des corrections de tests multiples (Bonferroni, Benjamini) peuvent être faites afin de renforcer la valeur des résultats statistiques. Les calculs de permutations nécessitent l'évaluation des scores d'enrichissement de listes de gènes obtenues par sélection aléatoire. La p-valeur ainsi obtenue reflète la probabilité d'occurrence de cette liste de gènes par rapport au hasard.

Une fois établies les listes de gènes différentiellement exprimés dans des conditions biologiques sélectionnés peuvent être analysées par regroupement. On pourra alors identifier des groupes de gènes significativement liés à des processus biologiques impliqués dans la problématique et le modèle étudié. Afin de valider expérimentalement par une autre approche le niveau d'expression de ces gènes candidats, on aura généralement recours à une expérience

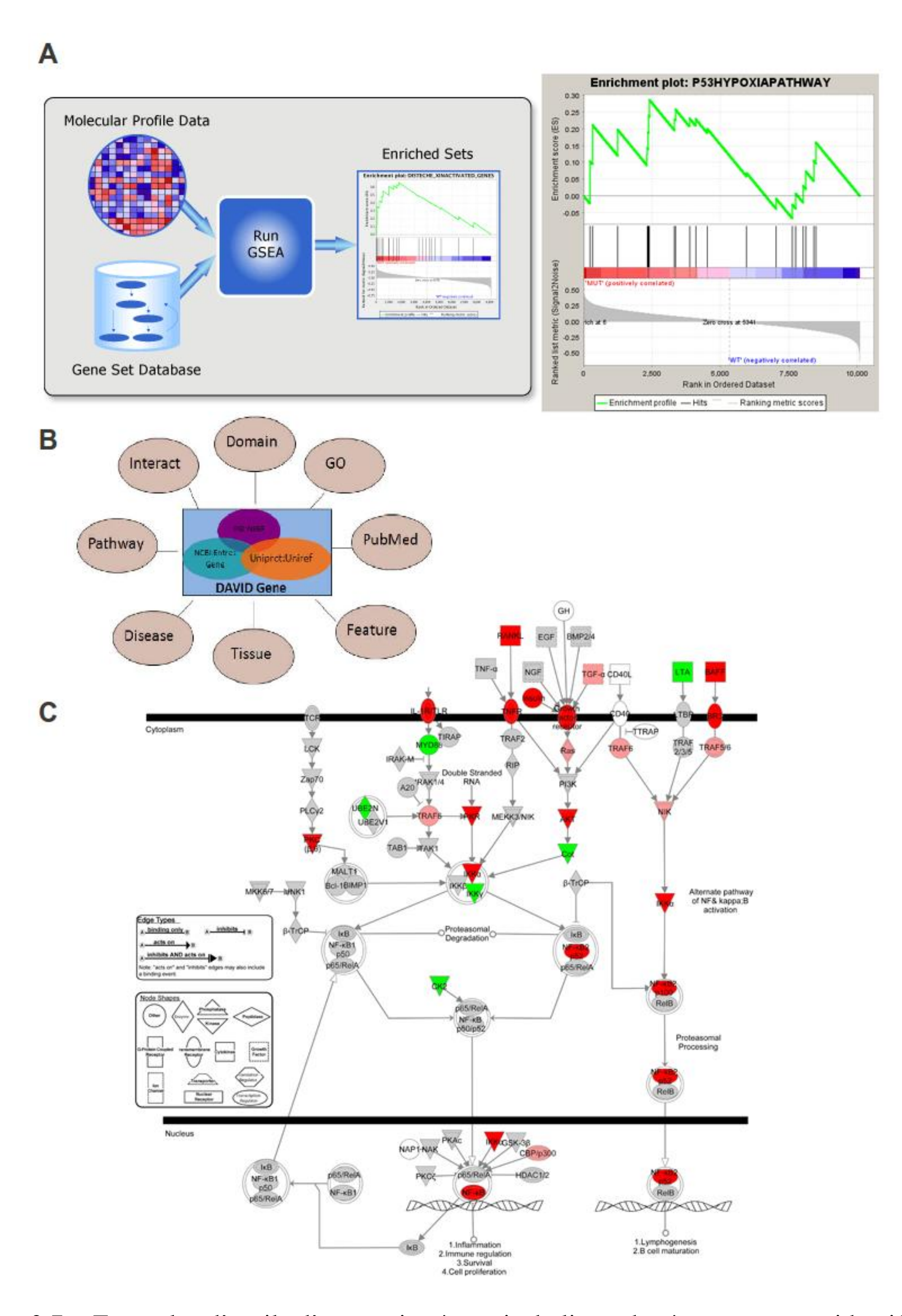


FIGURE 3.7 – Exemples d'outils d'annotation à partir de listes de gènes ou autres identifiants : (A) *Gene Set Enrichment Analysis* (GSEA), (B) *DAVIDknowledgebase* et (C) *Ingenuity Pathway Analysis* (IPA)

de PCR quantitative appelée couramment qRT-PCR.

D'un point de vue fonctionnel et analytique, il est ensuite possible et recommandé de contextualiser afin de générer des réseaux de gènes. La construction de tels réseaux permettra de comprendre et de modéliser les liens fonctionnels unissant les gènes discriminants précédemment identifiés. A terme, cela contribuera au décryptage du mécanisme de la pathologie ou du modèle étudié. De nombreux outils permettent de créer ce genre de réseaux (IPA, cytoscape).

3.4 Analyses de données dans le cadre de collaborations

Avant l'introduction de la technique de puces à ADN sur lame de verre au sein de la plateforme Transcriptome, le TAGC utilisait des puces radioactives sur support nylon, technique pour laquelle le laboratoire était précurseur. La plateforme TGML avait le matériel et les compétences requises pour réaliser la conception et le dépôt des sondes sur les puces, qui étaient ensuite utilisées dans diverses études. Ainsi, le développement de la librairie R AgiND (voir Chapitre 2) a donné lieu à de nombreuses collaborations, pour les deux technologies : puces à fluorescence Agilent™et puces radioactives nylon. Deux de ces collaborations, dans des domaines très différents, se sont concrétisées par des publications :

(1) la collaboration avec le Dr. Patricia Paris de l'IMTSSA (Institut de Médecine Tropicale du Service de Santé des Armées, Marseille) porte sur la définition d'une signature transcriptionnelle caractéristique d'infections virales provoquant la dengue. Elle à été réalisée avec des puces pan-génomiques Agilent™ ;

(2) la collaboration avec le Dr. El Chérif Ibrahim du NICN CNRS UMR 6184 (Faculté de Médecine Nord, Marseille) porte sur la caractérisation transcriptionnelle des voies de signalisation altérées lors du mauvais épissage alternatif du gène IKBKAP chez les patients souffrant de dysautonomie familiale. Ce projet a été mené en deux temps. Une première campagne sur puces à ADN de type nylon conçue entièrement sur la plateforme, a mis en évidence des gènes différentiellement exprimés entre les patients malades et les témoins. La seconde, sur puces à ADN pan-génomiques commerciales de technologie Agilent™, a permis de confirmer les précédents résultats, et d'étudier l'effet d'une molécule à visée thérapeutique.

Dans ces collaborations, mon travail a consisté à générer des signatures transcriptionnelles à partir des données brutes provenant des fichiers issus d'*Agilent Feature Extraction* (AFE) pour les puces Agilent™, du scanner pour les puces nylon. J'ai également formé et aidé les biologistes à utiliser les divers outils bioinformatiques : AgiND, TmeV, cluster, Treeview, DAVID knowledge database, *Ingenuity Pathways Analysis* (IPA), et contribué à la rédaction des articles.

3.4.1 La dengue

La dengue est une infection virale, endémique dans les pays tropicaux. Cette maladie est transmise à l'être humain par piqûres de moustiques Aedes aegypti infecté par un virus de la famille des flavivirus. De rares cas de formes asymptomatiques existent. Mais en général, cette infection virale entraîne des fièvres, maux de tête, douleurs musculaires et articulaires, fatigues, nausées, vomissements et éruptions cutanées. La fièvre peut être hémorragique avec ou sans syndrome de choc. Ce dernier cas est rare mais sévère et peut entraîner la mort du patient. L'OMS a défini en 1997 une classification clinique pour différencier les 3 principaux groupes de malades atteints de la dengue, mais celle-ci reste incomplète (Figure 3.8). Cette classification comprend : (1) la dengue classique (DF, « *Dengue Fever* »), (2) la dengue hémorragique sans syndrome de choc (DHF, « *Dengue Haemorrhagic Fever* ») et (3) la dengue hémorragique avec syndrome de choc (DSS, « *Dengue shock syndrome* »).

Nous avons utilisé une approche transcriptomique afin d'obtenir un aperçu des mécanismes moléculaires associés au développement de l'infection par la dengue avec syndrome de choc (DSS). L'objectif était, à terme, d'identifier des biomarqueurs de diagnostic de cette forme clinique pouvant être testés rapidement dans les pays endémiques, afin de pouvoir réduire le nombre de décès dus à cette maladie. Nous avons donc réalisé une analyse comparative des profils d'expression de cellules sanguines à partir des cellules de sang total (PAXgene blood RNA, Qiagen) chez 48 jeunes patients cambodgiens recrutés prospectivement pendant l'épidémie de dengue de 2007 et présentant des évolutions cliniques distinctes, selon la distribution suivante : DF (n=16), DHF (n=13) et DSS (n=19).

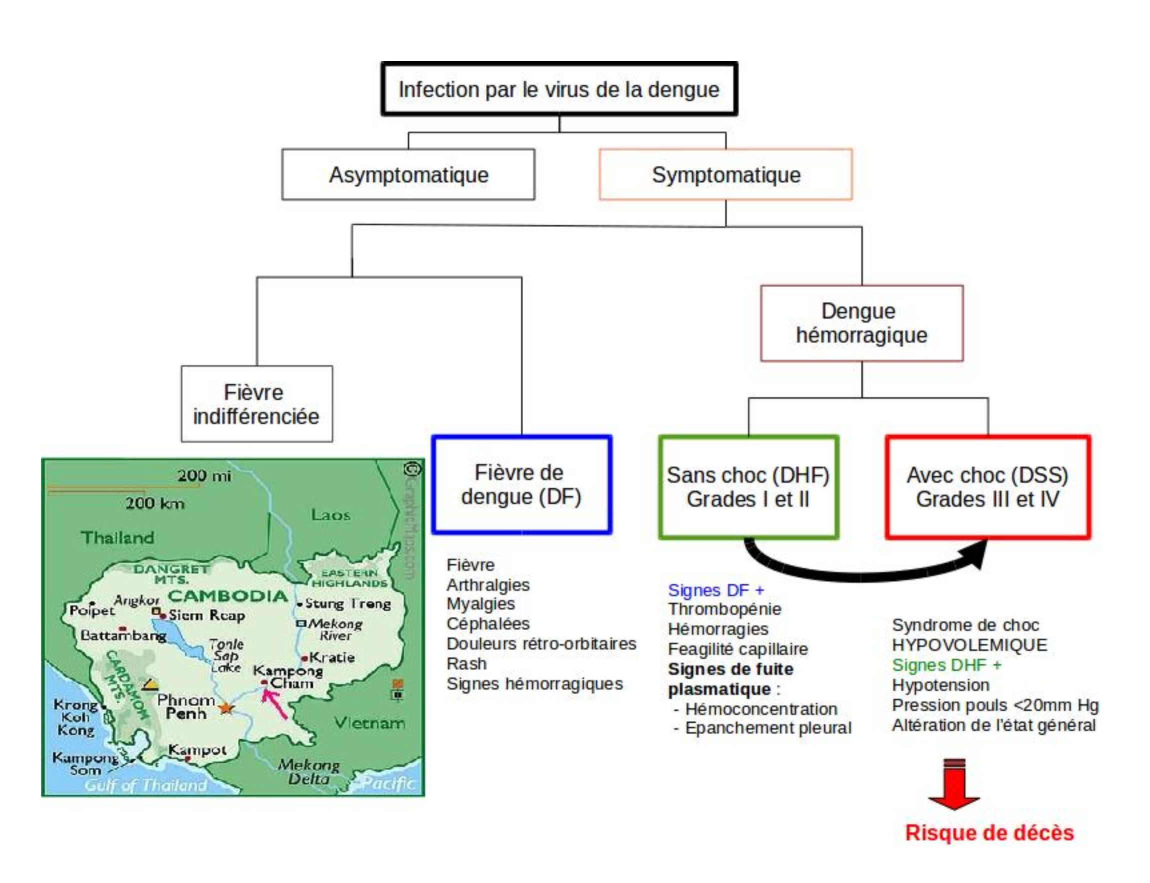


FIGURE 3.8 – Classification clinique de la dengue établie en 1997 par l'OMS et localisation de la région d'où proviennent les jeunes patients cambodgiens

OPEN ACCESS Freely available online



Genome-Wide Expression Profiling Deciphers Host Responses Altered during Dengue Shock Syndrome and Reveals the Role of Innate Immunity in Severe Dengue

Stéphanie Devignot[1], Cédric Sapet[1], Veasna Duong[2], Aurélie Bergon[3], Pascal Rihet[3], Sivuth Ong[2], Patrich T. Lorn[4], Norith Chroeung[4], Sina Ngeav[4], Hugues J. Tolou[1], Philippe Buchy[2], Patricia Couissinier-Paris[1]*

1 French Army Biomedical Research Institute (Institut de recherche biomédicale des armées, IRBA), Antenne de Marseille-IMTSSA; Unité de Virologie, Marseille, France, **2** Institut Pasteur in Cambodia, Department of Virology, Phnom Penh, Cambodia, **3** TAGC-INSERM U928, Marseille, France, **4** Kampong Cham Provincial Hospital, Kampong Cham, Cambodia

Abstract

Background: Deciphering host responses contributing to dengue shock syndrome (DSS), the life-threatening form of acute viral dengue infections, is required to improve both the differential prognosis and the treatments provided to DSS patients, a challenge for clinicians.

Methodology/Principal Findings: Based on a prospective study, we analyzed the genome-wide expression profiles of whole blood cells from 48 matched Cambodian children: 19 progressed to DSS while 16 and 13 presented respectively classical dengue fever (DF) or dengue hemorrhagic fever grades I/II (DHF). Using multi-way analysis of variance (ANOVA) and adjustment of p-values to control the False Discovery Rate (FDR<10%), we identified a signature of 2959 genes differentiating DSS patients from both DF and DHF, and showed a strong association of this DSS-gene signature with the dengue disease phenotype. Using a combined approach to analyse the molecular patterns associated with the DSS-gene signature, we provide an integrative overview of the transcriptional responses altered in DSS children. In particular, we show that the transcriptome of DSS children blood cells is characterized by a decreased abundance of transcripts related to T and NK lymphocyte responses and by an increased abundance of anti-inflammatory and repair/remodeling transcripts. We also show that unexpected pro-inflammatory gene patterns at the interface between innate immunity, inflammation and host lipid metabolism, known to play pathogenic roles in acute and chronic inflammatory diseases associated with systemic vascular dysfunction, are transcriptionnally active in the blood cells of DSS children.

Conclusions/Significance: We provide a global while non exhaustive overview of the molecular mechanisms altered in of DSS children and suggest how they may interact to lead to final vascular homeostasis breakdown. We suggest that some mechanisms identified should be considered putative therapeutic targets or biomarkers of progression to DSS.

Citation: Devignot S, Sapet C, Duong V, Bergon A, Rihet P, et al. (2010) Genome-Wide Expression Profiling Deciphers Host Responses Altered during Dengue Shock Syndrome and Reveals the Role of Innate Immunity in Severe Dengue. PLoS ONE 5(7): e11671. doi:10.1371/journal.pone.0011671

Editor: Patricia T. Bozza, Fundação Oswaldo Cruz, Brazil

Received January 20, 2010; **Accepted** June 22, 2010; **Published** July 20, 2010

Funding: This work was supported by a clinical research program PRC 2007–13 from the French Army Medical Health Service (Service de Santé des Armées Françaises). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: parisp@imtssa.fr

These authors contributed equally to this work.

Introduction

Acute dengue virus infections are a major public health problem for many tropical and sub-tropical countries and an increasing risk for the worldwide population [1]. Symptomatic infections occur under a spectrum of diseases ranging from classical dengue fever (DF) to the most severe life-threatening dengue shock syndrome (DSS), a leading cause of childhood hospitalisation and death in endemic countries with limited health resources [1,2].

DSS is regarded as a vascular disease involving a complex interplay between virus, whole blood cells and microvascular territories [3,4], and thought to result largely from an aberrant host response to infection. As for other major systemic diseases, a detrimental cytokine storm is thought to be central to the systemic microcirculatory failure and massive plasma leakage leading to cardiovascular decompensation characterizing DSS [5]. However, controversies exist regarding the nature of pathogenic host immune responses supporting this life-threatening syndrome [6–8]. Indeed, reactivation of cross-reactive memory T lymphocytes and increased infection of monocytes mediated by cross-reactive antibodies acquired during previous infections by distinct dengue virus serotypes, are the main hypothetic mechanisms proposed to explain the putative cytokine storm leading to plasma leakage [5,9]. However, those hypothesis fail to explain the occurrence of DSS in patients having primary dengue infection and their relevance to the pathophysiology of DSS disease is discussed [8,10].

Efforts to identify soluble biomarkers of severe dengue differentiating uncomplicated dengue infections from severe ones has led to the identification of a diversity of cytokines, chemokines, endothelial agonists or soluble endothelial molecules [11–18]. However, discrepancies in definition of dengue severity, variability in patients cohorts characteristics, as well as in techniques and markers investigated, have impaired the identification of reliable sets of DSS biomarkers and the possibility to get a global overview of biological markers altered during DSS.

Understanding the molecular basis of DSS and identifying relevant DSS biomarkers thus remains a major challenge [5,6]. Indeed, DSS occurs by the end of the acute infection in only a fraction of dengue-infected patients and current severity criteria, based on the 1997 World Health Organization (WHO) classification of dengue severity, fail to predict a significant proportion of patients who progress to life-threatening DSS [19–21]. Attempting to decipher molecular mechanisms underlying DSS by analyzing circulating whole blood cell genome-wide expression profiles is a relevant approach regarding the study of other systemic inflammatory syndromes, where a cognate cross-talk between endothelial vascular cells and blood cells occurs [22–24]. Whole blood represent a highly informative while complex cellular sample, that may reflect host pathophysiological responses ongoing at the time of blood sampling [22]. Furthermore, whole blood cells are easy to collect and store during field studies on large cohorts, reducing samples volumes required and limiting technical bias due to cell purification. However, due to the high cellular complexity of whole blood cells samples, whole gene expression patterns should be carefully analyzed and deciphered to allow returning to an integrative view of the molecular mechanisms altered during the pathophysiological process studied [25].

Such a bench-to-bedside medical research has gained more and more interest in the recent years. Indeed, it allowed improving the understanding of pathophysiological processes underlying systemic critical illnesses such as sterile and non sterile systemic inflammatory responses syndromes (SIRS), allowing the identification of relevant disease biomarkers and of new putative therapeutic targets [22–24,26].

Genome-wide expression studies aimed at deciphering molecular responses altered in the whole blood cells of adults [27] and children DSS patients [28] have been implemented recently by colleagues. They reported a decreased IFN type I-induced response and a benign transcriptional response at the time of cardiovascular decompensation [27,28], but failed in identifying biological pathways relevant to DSS pathophysiology and particularly inflammatory ones that could sustain microvascular dysfunction [28].

We report here the results of a prospective study comparing the whole blood genome-wide expression profiles of 48 matched Cambodian children recruited during the huge 2007 dengue outbreak who presented with classical dengue fever (DF), dengue hemorrhagic fever grades I/II (DHF) or dengue shock syndrome (DSS), according to the 1997 WHO classification of dengue severity [29]. Based on careful study design and statistic treatment of microarrays data, we identified a large and highly relevant gene signature of DSS never reported before, that discriminates DSS children from paediatric patients with DF or DHF grades I/II, who did not present severe clinical complications. Using an integrative analysis of the gene patterns altered in DSS children, we deciphered part of the complex interactive molecular processes occurring during DSS, highlighting similarities between DSS and other major inflammatory processes. Finally, we identified unexpected pro-inflammatory innate immune responses activated in the whole blood cells of DSS children that may play a major role in DSS pathophysiology. The implications of present findings to the improvement of DSS prognosis and treatment are discussed.

Materials and Methods

Ethics statement

The global study and all protocols presented here were approved by the national Cambodian ethical committee. Written informed consent was obtained from the legal guardians of each child. To ensure strict anonymity regarding the patients, samples were encoded as PLxxx (Plasma Leakage).

Patients and clinical data

Inclusion criteria retained were: age (1 to 15 years old); positive diagnosis of acute dengue infection assessed by different methods; absence of known chronic inflammatory disease or ongoing acute co-infection at the time of inclusion.

An eligible cohort of 83 dengue-infected children hospitalised at the Kampong Cham provincial hospital, Cambodia, was prospectively enrolled from July to September 2007 during the huge 2007 dengue outbreak in Cambodia, characterized by a high number of DSS cases.

Children diagnosed with acute dengue infections were classified at admission as classical dengue fever (DF), dengue hemorrhagic fever (DHF) or dengue shock syndrome (DSS) based on the 1997 WHO criteria [29]. Clinical and biological follow-up was done daily for each hospitalised patient. DSS patients were admitted to hospital intensive care unit where they received appropriate fluid resuscitation and were monitored for vital parameters. Children who required blood transfusion were not included in the study.

To increase the probability to identify gene signatures specific of DSS, we chose to include only symptomatic dengue-infected classified DF, DHF and DSS, but no healthy or non-dengue children in the present study. This is based on the rationale that comparing DF, DHF and DSS patients together should improve the probability to identify a DSS-specific gene signature, while including an external non dengue control group should increase the probability to identify a general dengue-related signature but should be less powerful at identifying a signature of severe dengue disease.

DF, DHF and DSS patients whole blood samples selected for the present study corresponded to comparable duration of illness after onset of fever: all were collected within a window of time comprised between 3 days and 7 days after onset of fever (being considered day 0). For most DSS patients, this generally corresponded to the day of cardiovascular decompensation (shock) or the day after, except for 3 (PL017, PL033, PL047) and 2 DSS (PL005, PL101) for whom blood was collected respectively 2 and 3 days after onset of shock.

Patients' samples selected for the present study were also carefully matched for age, gender, viral serotype (when identified) and immunological status (primary or secondary, according to reference assays described in diagnosis methods) towards dengue infection.

Diagnosis assays carried out as described thereafter, indicated that about 90% of all dengue-infected children had secondary infection.

Dengue diagnosis and immunological status

All diagnosis assays were carried out at the Institut Pasteur in Cambodia, the National Reference Center for arboviral diseases in Cambodia. IgM capture ELISA and Hemagglutination-inhibition were performed on paired sera collected at admission and at discharge, and systematically tested for both dengue and Japanese

Encephalitis virus, another flavivirus endemic in Cambodia, as described previously [30]. Virus isolation was carried out on earliest serum samples by inoculating permissive C6/36 and VERO E6 cells, followed by serotype-specific immunofluorescence [30]. Viral RNA was detected in specimens collected at early stage of the disease using a nested RT-PCR [31]. Primary or anamnestic "secondary" antibody response indicating previous infections by dengue viruses was determined from paired serum samples by hemagglutination-inhibition assay. Interpretation followed WHO recommendations [29].

Blood sample preparation

Whole blood samples (2.5 ml) were collected on PAXgene^{TM} Tubes (PreAnalytiX^{TM}) further stored at −80°C, before being sent to France in dry ice. Extraction of series of 24 matched samples (DF, DHF and DSS) was done using PAXgene^{TM} Blood RNA kits (PreAnalytiX^{TM}) rapidly after collection. Purified total RNAs kept at −80°C were processed for hybridization on genome-wide DNA microarrays within one month.

cRNAs preparation and microarrays hybridization

All RNAs were checked for integrity using the 2100 BioAnalyzer (Agilent Technologies) and quantified using a ND-1000 spectrophotometer (NanoDrop Technologies). Cyanine-3-labeled cRNA was generated from 0.3 µg of RNA using the One-Color Low RNA Input Linear Amplification kit (Agilent) according to the manufacturer's instructions, followed by purification on RNAeasy column (QIAGEN). All amplified cRNAs were checked for dye incorporation, cRNA yield and amplification profile. Only those fitting all quality criteria were fragmented for further hybridization on microarrays. Samples from DF, DHF and DSS patients were then carefully matched and hybridized onto Agilent Whole Human Genome (4×44K) Oligo Microarrays (G4112F). Microarrays were scanned using an Agilent DNA microarray scanner G2505B.

Microarray data analysis

All microarray data is MIAME compliant and the raw and normalized data have been deposited in the MIAME compliant database Gene Expression Omnibus [32] (GEO Series accession number GSE17924 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = GSE17924).

Individual microarray quality was evaluated based on QC report, pair-wise MA-plots, and box plots. Intra-array normalization of raw signals from the 48 microarrays was done using Feature Extraction software 9.1.3.1 (Agilent). Microarrays normalized data were further exported into the Limma package [33], for inter-array normalization using the quantile method [34].

Statistical analysis were was performed using the TIGR MeV (MultiExperiment Viewer) v 4.4 software (http://www.tm4.org/mev.html), and the GeneANOVA program [35]. Multi-way ANOVA model was implemented: first, to identify differentially regulated genes when accounting for the multiple sources of variation in the microrray experiment; second, to evaluate the effect of the main variable, disease phenotype, relatively to that of other putative confounding variables such age, gender, duration of illness or microarray technical variability (independent extractions or hybridizations). Local ANOVA further determined the contribution of each covariate on the expression level of each gene. Multiple test correction was further carried out using the false discovery rate (FDR) method [36]. Cluster [37] and Tree View softwares [38] were used for unsupervised hierarchical clustering. Iterative SVM (Support Vector Machine) method associated with leave-one-out cross-validation [39] was used to assess the robustness of DSS-gene signature.

Real-time PCR validation of genes over and under-expressed in DSS patients

Briefly, total RNA extracted from whole blood samples was reverse-transcripted using the High Capacity cDNA RT kit (Applied Biosystems Inc) and random primers. Real-time PCR were carried out using the FastStart Universal Probe Master (ROX) (Roche) and real-time PCR primers designed using the Universal Probe Library (UPL) Assay Design Center (Roche). Amplification products were run on an ABI-PRISM 7900HT (Applied Biosystems). Cycle threshold Ct values were automatically calculated and value obtained for each gene amplified was normalized by subtracting the Ct corresponding to amplification of the HPRT1 gene (ΔCt) for the same sample. Correlation between ΔCt values obtained by real-time PCR and corresponding expression values from microarrays was estimated using Spearman correlation coefficient.

Comprehensive overview of functional patterns altered during DSS

Bio-informatics-based analysis using the demonstration version 7.1 of Ingenuity Pathway Analysis software (IPA; Ingenuity® Systems, www.ingenuity.com) associated with manual and litterature-based analysis was carried out to identify the most relevant functional processes associated with the identified DSS gene signature. This was done by combining most informative canonical pathways identified using IPA, genes having the strongest association with the disease phenotype based on ANOVA analysis, and similarities to molecular patterns altered in other systemic inflammatory processes associated with endothelial dysfunction.

Results

Patient characteristics

To identify gene patterns specifically altered in DSS patients, we compared three groups of carefully matched paediatric patients representing the main clinical forms of symptomatic dengue infections DF (n = 16), DHF (n = 13) and DSS (n = 19), according to the 1997 WHO classification criteria of dengue severity [29]. Altogether, DF, DHF and DSS represent different subtypes of the disease phenotype variable, further considered in this study. The clinical characteristics and values of haematological parameters are presented in table 1 (median values from each patient group) and table S1 (individual values from each of the 48 patients included). Supportive treatment provided to DSS patients are mentioned. As indicated, DSS children had significant lower relative neutrophil counts (median values / DF: 3900; DHF: 3950; DSS: 2500; p-value: 0.03; Kruskal Wallis test).

Unsupervised hierarchical clustering discriminates DSS children from DF/DHF ones, revealing a DSS-gene signature

Since microarray data analysis can be affected by a number of bias [40], we put a particular care on study design and analysed data from the 48 normalized microarrays using multi-way analysis of variance (ANOVA) [41]. Indeed ANOVA evaluates the statistical probability (p-value), for each individual gene, that a difference in expression between the three patient groups could have been observed by chance. This allows revealing genes that show even small but highly significant changes in expression, regarding the studied phenotype.

Table 1. Clinical and biological characteristics of DF, DHF and DSS patient groups at the time of hospital admission.

	DF ($n=16$)	DHF ($n=13$)	DSS ($n=19$)
Patients characteristics			
gender, male *n* (%)	7 (43%)	4 (31%)	7 (37%)
age, median (IQR), years	8 (4–9)	7 (5–8)	8 (7–9)
weight, median (IQR), kg	18 (13–20)	15 (14–18)	19 (15–23)
hospital admission, median (IQR), day after onset of fever (D0)	2 (1–3)	2 (2–3)	4 (3–4)
Dengue status			
viral serotype, *n* (DENV-1/DENV-2/DENV-3/DENV-4/unknown)	4/2/8/1/1	1/1/10/1/0	1/1/10/0/7
immunological status, secondary infections, *n* (%)	14 (88%)	12 (92%)	18 (95%)
Clinical manifestations			
tourniquet test (pos/neg/not done) (%)	56%/44%/0%	54%/38%/8%	37%/32%/31%
hepatomegaly, *n* (%)	3 (19%)	6 (46%)	17 (89%)
gastro-intestinal bleeding, gingivorragy, hematemesis, melena, *n* (%)	0	1 (8%)	6 (32%)
Blood pressure			
heart frequency, median (IQR), pulse per minute	113 (100–124) ($n=14^a$)	120 (112–120)	Not perceptible ($n=15^a$)
pulse pressure, median (IQR), mm Hg	40 (30–45)	30 (30–40)	15 (10–20) ($n=15^a$)
Haematological parameters			
thrombocytopenia (platelet count $<100000/mm^3$), %	15% ($n=14^a$)	55% ($n=11^a$)	94% ($n=17^a$)
hematocrit, median (IQR), %	36.5 (35–39) ($n=14^a$)	39.75 (38–42) ($n=12^a$)	42.5 (38–45) ($n=18^a$)
hemoconcentration (hematocrit $>20\%$), *n* (%)	1 (6%)	3 (23%)	17 (89%)
white blood cells, median (IQR), number/mm^3	6600 (5500–9900) ($n=13^a$)	6450 (6200–7400) ($n=10^a$)	6900 (4800–6900) ($n=17^a$)
neutrophils, median (IQR), number/mm^3	3900 (2900–7600) ($n=13^a$)	3950 (3500–4200) ($n=10^a$)	2500 (2200–3800)($n=17^a$)
lymphocytes, median (IQR), number/mm^3	1600 (1400–2100) ($n=13^a$)	1850 (1500–1900) ($n=10^a$)	2200 (1500–3200) ($n=17^a$)
Supportive medical care			
oxygen supplementation, *n* (%)	0	0	15 (79%)
perfusion of colloid (dextran 40), *n* (%)	0	0	14 (74%)
perfusion of human plasma, *n* (%)	0	0	8 (42%)

DENV, dengue virus; DF, dengue fever; DHF, dengue hemorrhagic fever; DSS, dengue shock syndrome; IQR, interquartile range; n, number.
[a]n = x : with x : number of patients for which the data is available.
doi:10.1371/journal.pone.0011671.t001

Based on ANOVA analysis, lists of genes differentially expressed between DF, DHF and DSS groups were generated using different false discovery rate (FDR) ranging from 0.05 up to 10%. Indeed, low FDR provide more stringent statistical filter while they reduce the number and thus the enrichment of genes differentially expressed. At the opposite, higher FDR, while statistically accepting a higher number of false positive genes, also provide larger and enriched gene lists that should be more informative when searching to identify molecular pathways. Based on this rationale, we chose to work using the gene list generated at FDR 10 after we verified by a different statistical method currently used for the analysis of microarrays data, SAM (Significant Analysis of Microarray) [33], that most significant genes were commonly found by the two types of analyses (data not shown). The gene list generated at FDR10 included 2959 genes differentially expressed between DF, DHF and DSS patients groups (Table S2).

The biological relevance of those differentially expressed genes was assessed using local ANOVA that allows evaluating the contribution of the main variable, disease phenotype, and that of other putative confounding variables related to patients (age, gender, day of blood sampling, viral serotype) and to technical steps (effect of independent RNA extractions, amplifications and hybridization) on variations of expression levels of those 2959 genes. This confirmed that the disease phenotype strongly influenced the variations of expression of the 2959 genes differentially expressed between the three patient groups, reinforcing the biological significance of this set of genes (Table S2).

Unsupervised hierarchical clustering based on the 2959 gene signature identified was then applied to the 48 children gene expression profiles. This allows clustering the patients whose gene expression profiles are the more similar independently of their disease phenotype subtype. As a result, the 48 patients expression profiles were organized in two major subsets (Figure 1): subset 1 (first dendrogram branch) includes both DF and DHF patients without distinction; subset 2 (second dendrogram branch) encompasses a sub-group (2a) of DF and DHF patients, and a distinct sub-group (2b) including 17 out of the 19 DSS patients, whatever they received or not plasma infusion, revealing a "DSS-gene signature" common to most DSS patients. Some few patients clustered however in unexpected subsets: two DSS patients (PL005, PL101) had gene expression profiles closer to those of the DF/DHF 2a subset, while one DF patient (PL064) and three DHF patients (PL037, PL058, PL070) gene expression profiles clustered within the DSS 2b subset.

We confirmed the robustness of the DSS-gene signature using the iterative Support Vector Machine (SVM) classifier learning method [39], which reclassified all the 19 DSS patients together.

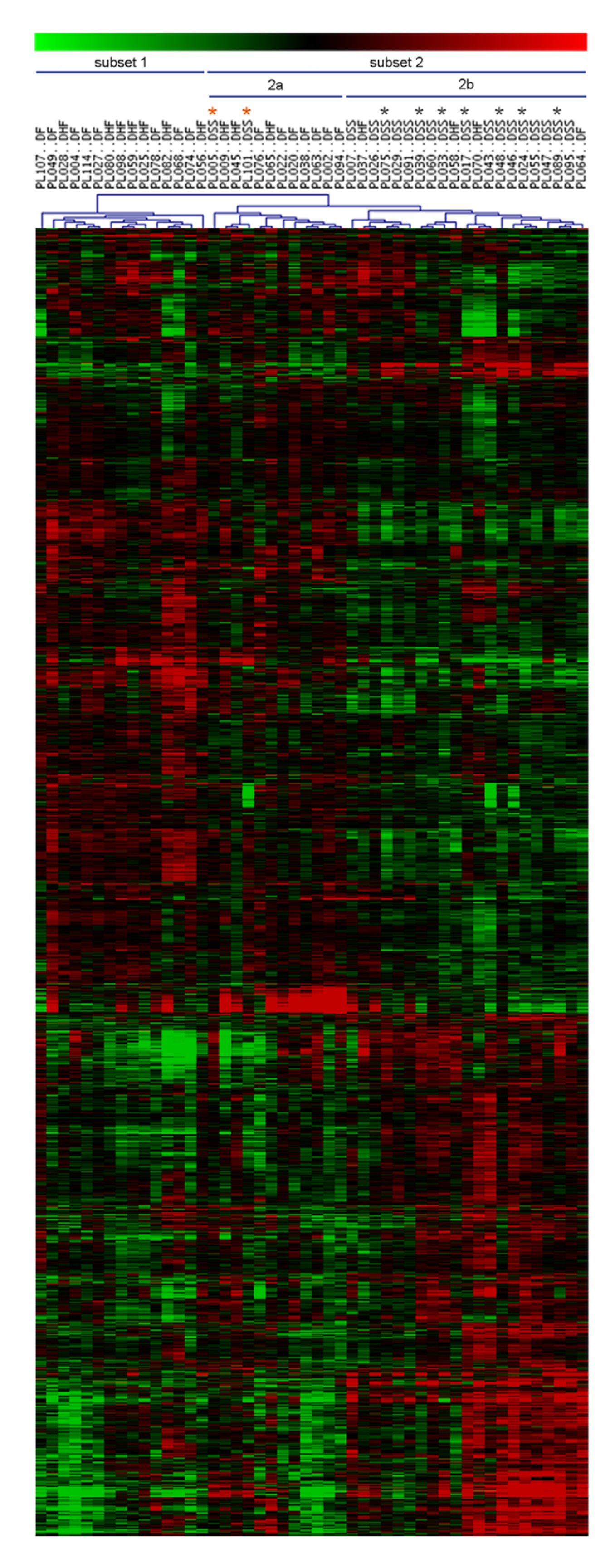


Figure 1. Unsupervised hierarchical clustering of whole blood cells expression profiles from the 48 dengue-infected children. The clustering is based on the 2959 gene list (3515 clones, detailed in Table S2) discriminating dengue fever (DF), dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS) patients. Each row represents a single transcript and each column represents a patient's sample. Color scale indicates the range of gene expression: black indicates median expression level, red greater expression, green lower expression. The 2 patient subsets identified are indicated. PLxxx, code relative to one patient. Black star: DSS patient sampled 3 days after shock. Orange star: Patients who received perfusion of human plasma before collection of blood samples.
doi:10.1371/journal.pone.0011671.g001

To validate microarray data, we carried out real-time RT-PCR focusing on nine genes strongly associated with the DSS-gene signature, using 15 patients samples (five from each disease phenotype subtype: DF, DHF and DSS). Results obtained strongly correlate microarray data (Figure S1).

DSS-gene signature analysis identifies a diversity of genes and canonical molecular pathways related to immunity, inflammation and host metabolism

Filtering genes from those having the highest to the lowest statistical association with the disease phenotype variable (Table S2), relying on results from multi-way and local ANOVA revealed that the individual genes having the strongest association with the DSS phenotype subtype are, for a large part, related to innate immunity, inflammation and host lipid metabolism, a finding confirmed when the whole 2959 genes of the DSS-gene signature were processed through the knowledge-based IPA software.

Indeed, IPA analysis identified that 163 canonical pathways were significantly associated with those genes (data not shown) with a large proportion of immune-related pathways in the first top 30 (Figure 2). In particular, several under-expressed but partially redundant signaling canonical pathways related to T lymphocyte activation were identified, of which the T cell receptor (TCR) signaling pathway (Figure 3), which has the strongest association with the DSS-gene signature. Interestingly, a number of metabolic pathways, and particularly of lipid-signaling pathways, were significantly represented among the 163 DSS-related canonical pathways.

When comparing our results to those of colleagues who reported gene or protein signatures associated with DSS, we identified some transcripts encoding proteins considered putative markers of severe dengue. This includes non exhaustively the acute phase pentraxin-related protein PTX3 [15], the anti-inflammatory IL-10 [11] or the pro-inflammatory IL-18 [12] cytokine transcripts that have increased abundance in the DSS-gene signature, while having intermediate to low statistical association with the disease phenotype variable according to the multi-way ANOVA (Table S2).

IFN type I-related transcripts, of which abundance was shown to be decreased in DSS patients by others [27,28,42], represented only a limited number of genes associated with the DSS-gene signature. This was confirmed by the IPA analysis that did not identify IFN type I-related pathways among those strongly associated with the DSS-gene signature (Figure 2).

DSS is associated with impaired expression of T and NK cell-related genes but increased expression of anti-inflammatory and repair/remodeling transcriptional responses

Integrative analysis of the most significant individual genes and canonical pathways extended the finding that a large and diverse set of genes related to T but also to NK lymphocyte activity is

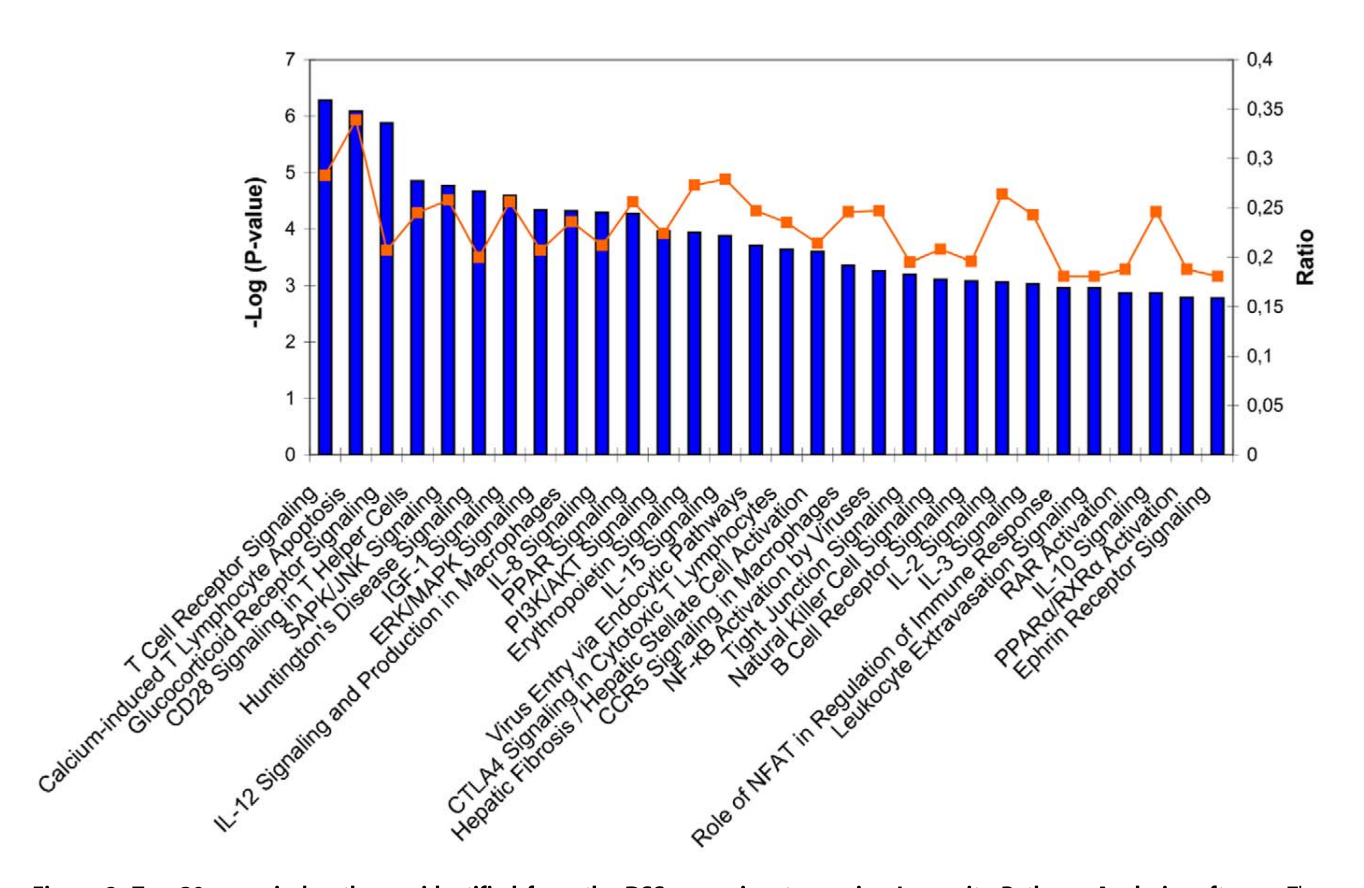


Figure 2. Top 30 canonical pathways identified from the DSS-gene signature using Ingenuity Pathway Analysis software. The significance of the association between data set and canonical pathway was estimated by the p-value (Fischer's exact test; left axis) and the ratio (right axis) of genes that maps to each canonical pathway.
doi:10.1371/journal.pone.0011671.g002

under-expressed in DSS patients compared to DF and DHF counterparts (Table 2; non exhaustive list; individual p-values available in Table S2; and Figure 3). Those genes are critical to a number of T and NK-cell functions, including T and NK cell differentiation, receptor signaling, activation and proliferation, cytotoxic functions or recruitment of lymphocytes to peripheral tissues. Since lymphocyte counts did not differ between the DF, DHF and DSS children ($p = 0.428$; Kruskal Wallis test), we searched whether genes encoding factors regulating negatively T and NK functions were over-expressed in the DSS-gene signature. We identified that the two genes having the strongest association with the disease phenotype variable, encode two major immunomodulatory factors, the microsomal prostaglandin E synthase (*PTGES*/Agilent clone number A-24-P478940) and the complement regulatory protein CRIg/VSIG4, considered potent negative regulators of T and NK cell responses [43–45]. The decreased abundance of *NFkB* signal transduction-related transcripts (Table 2), already reported in DSS patients by others [46], might be related to impaired expression of T and NK cell-related genes.

Our analysis also revealed that DSS whole blood cells from children over-expressed an enriched pattern of anti-inflammatory and repair/tissue remodeling genes (Table 3; non exhaustive list; individual p-values available in Table S2). Over-expressed anti-inflammatory genes identified encode molecules with diverse functions: the anti-inflammatory cytokine IL-10, a putative marker of severe dengue [11], serine proteases and metalloprotease inhibitors, IL-1β cytokine decoy receptor, free heme scavenger molecules, or complement-regulating receptors. Repair and remodeling genes over-expressed in the DSS-gene signature also encode a diversity of molecules: the MMP-9, a matrix metalloprotease with key role in tissue remodeling and a candidate to dengue plasma leakage [47], the extracellular matrix molecules fibronectin, versican and collagens, the angiogenin and VEGF [17] endothelial agonists as well as the arginase 1 repair enzyme, which competes with the endothelial NOS (*NOS3*) for L-arginine biodisponibility [48].

Thus, DSS children whole blood cells have a global decreased abundance of T and NK cell-related transcripts but an increased abundance of anti-inflammatory and repair/remodeling transcripts at the time of cardiovascular decompensation.

Pro-inflammatory innate defense and host-lipid metabolism-related transcriptional responses are activated in DSS children

When searching for pro-inflammatory gene patterns that may be relevant to DSS pathophysiology and particularly to systemic inflammation and vascular dysfunction, we identified three major pro-inflammatory gene patterns. Interestingly, all are related to innate defense and host lipid metabolism, and considered major pathogenic mechanisms in other systemic inflammatory diseases.

As shown in table 4 (non exhaustive list; individual p-values available in Table S2), the first one is defined by a set of over-expressed genes strongly associated with the disease phenotype subtype DSS that encode highly pro-inflammatory microbicidal peptides and enzymes. This pattern includes non exhaustively the alpha defensins DEFA1, DEFA3 and DEFA4, the cathelicidin (CAMP) and lactoferrin (LTF) peptides, the neutrophil enzymes

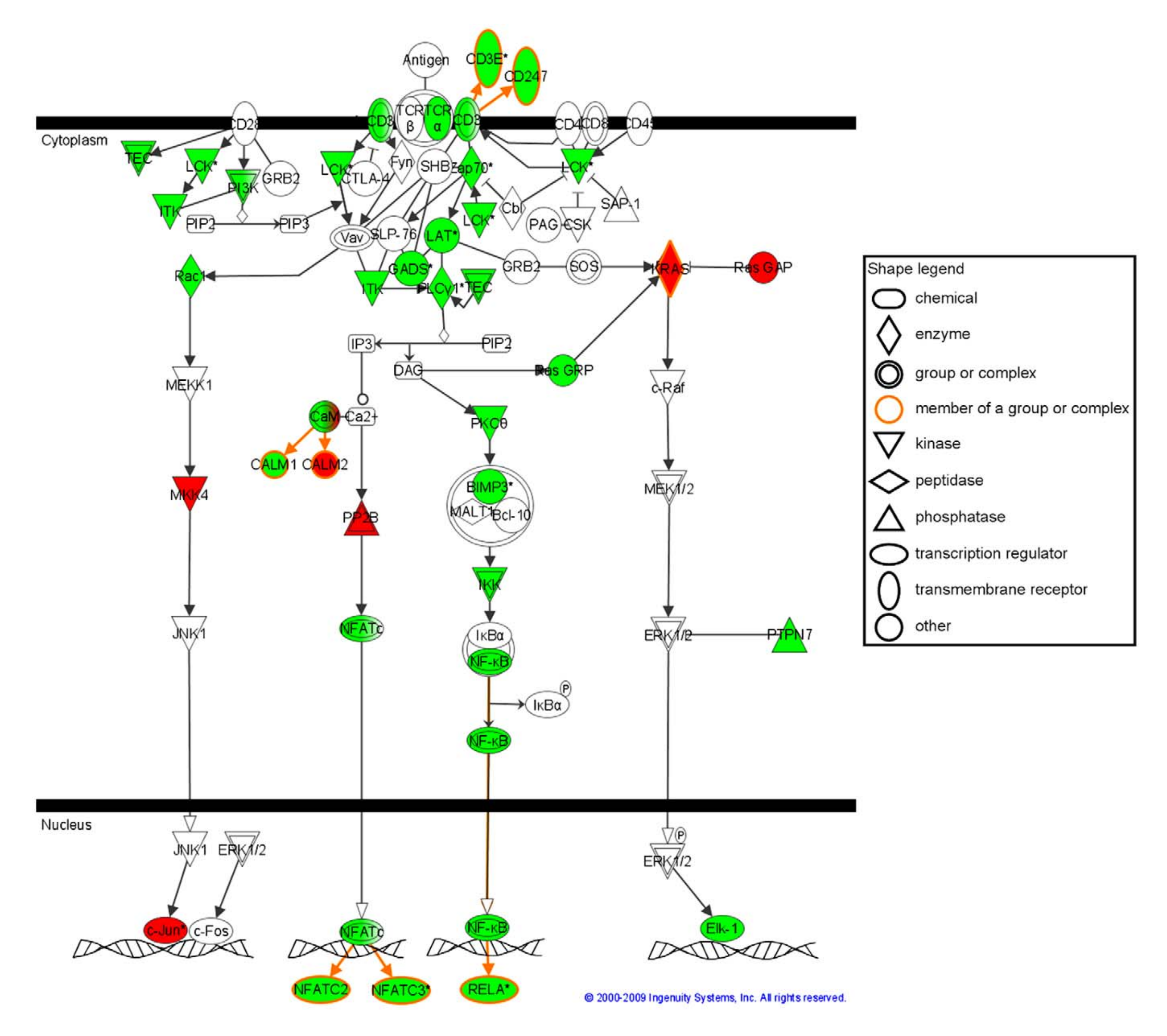


Figure 3. T Cell Receptor Signaling canonical pathway from Ingenuity Pathway Analysis. Genes in green and red are respectively under- and over-expressed in the DSS-gene signature. Genes in white are other genes present in the canonical pathway but absent from the DSS-gene signature. DSS: Dengue Shock Syndrome.
doi:10.1371/journal.pone.0011671.g003

myeloperoxydase (MPO), neutrophil RNASE2, RNASE3, cathepsin G and neutrophil elastase (ELANE). Transcripts encoding the potent pro-inflammatory calgranulins proteins S100A8/9 and S100A12, characteristic of granulocyte neutrophil activity [51] and involved in a diversity of inflammatory diseases [56], as well as the granulocyte-related metalloprotease MMP8 are also over-expressed. Increased abundance of those transcripts cannot be explained by increase in granulocyte count, since DSS patients have lower relative granulocyte counts than DF and DHF counterparts (median values/DF: 3900; DHF: 3950; DSS: 2500; p-value: 0.03; Kruskal Wallis test), thus reflecting more likely cellular activation. Altogether those results show that a transcriptional pattern of innate defense genes is activated in the whole blood of DSS children.

The second pro-inflammatory gene pattern identified is typical of altered homeostasis of cholesterol in monocytes/macrophages that characterises inflammatory lipid-laden monocytes/macrophages (lipid-laden Mo/Mac), a subtype of foam cells initiating vascular lesions in metabolic inflammatory diseases [57–59] (Table 5; non exhaustive list; individual p-values available in Table S2). Since the *PPARG* gene, which encodes a nuclear-lipid receptor involved in lipid signaling and lipid-homeostasis in inflammatory lipid-laden Mo/Mac [65], has a very strong association with the dengue disease phenotype, we searched whether other genes involved in cholesterol homeostasis in Mo/Mac had altered expression in the DSS-gene signature. Remarkably, we found a large lipid-laden Mo/Mac-related gene expression pattern characterized in particular by a higher abundance of transcripts encoding the key scavenger receptors of modified low density lipoproteins OLR-1, CD36 and MSR1, but a decreased abundance of transcripts encoding critical cholesterol transporters such as the NPC1 [66] or the ABCA1-

Table 2. T lymphocytes and NK cells-related genes present in the DSS-gene signature.

Function	Genes	P-value	Var.[a] (%)
Th1 differentiation	RUNX3, STAT4, TBX21	<0.00001 to 0.00242	25 to 42
Th2 differentiation	GATA3, STAT5A	0.00003 to 0.00225	17 to 32
Cytotoxic T lymphocyte functions	CTSW, PRF1	0.00005 to 0.00231	21 to 33
T lymphocyte activation	IL2RB, IL2RG	0.00014 to 0.00039	29 to 35
Cooperation with antigen-presenting cells	CD40LG	0.00105	21
Recruitment and interaction of T lymphocytes with endothelium	ITGAL, XCL1, XCL2	<0.00001 to 0.00214	20 to 33
Inhibitory NK cell receptors	KLRD1	0.00001	31
Activating NK cell receptors	NCR1, NCR3, CD160	<0.00001 to 0.00069	28 to 39
Cytotoxic molecules	GZMM	<0.00001	32
Receptors for NK cells homing to peripheral tissues	S1PR5	<0.00001	48
Differentiation factors of NK cells	FLT3LG, **IL15**, **IL17C**, KITLG	0.00088 to 0.00774	13 to 21
Suppression of T lymphocytes and NK cells response	**PTGES**, **VSIG4**	<0.00001	60 to 63
NFκB-related genes	**IRAK3**, TNIK, RELA, NFKBIB, TRAF1, TRAF2, TRAF6	0.00001 to 0.00506	9 to 31

HUGO gene names are indicated. When genes were represented by several clones on the microarray, p-value and variance medians were calculated. Genes in regular and bold are respectively under- and over-expressed in dengue shock syndrome patients.
[a]percentage of variance associated to disease phenotype.
doi:10.1371/journal.pone.0011671.t002

like ABCA10 [86], which regulate the efflux of modified cholesterol from Mo/Mac. Other lipid-laden cells-related genes also have altered expression in the DSS-gene signature. In particular, the *PPARA* gene that negatively regulates the formation of lipid-laden Mo/Mac [87] has decreased abundance in DSS patients. At the opposite, transcripts encoding the chitinase 1, a marker of pro-inflammatory lipid-laden Mo/Mac [73], and the FABP4, SOCS6, RETN and IRS2 proteins involved in lipid-laden Mo/Mac-induced insulin-resistance and compensatory response [68–72], have all increased abundance, also strongly supporting a biological signature of foam cells. Interestingly, the *PCSK9* transcript, which encodes a secreted protein that decreases the recycling of LDL to the liver by inducing the degradation of liver LDL receptors [74], is also over-expressed in the DSS signature and highly associated with the disease phenotype.

Thus, a gene expression pattern similar to that characterizing lipid-laden monocytes, is activated in the whole blood cells of DSS children at the time of cardiovascular decompensation.

The third pro-inflammatory gene pattern associated with the DSS-gene signature is characteristic of the metabolic pro-inflammatory arachidonic-acid pathway, one of the lipid metabolic pathways identified through IPA. As shown in table 5, the gene encoding the upstream cytosolic phospholipase

Table 3. Anti-inflammatory, tissue remodeling and repair genes present in the DSS-gene signature.

Function	Gene Symbol	P-value	Var.[a] (%)
Anti-inflammatory genes			
immunoregulatory molecules	**IL10**	0.00430	20
anti-proteases	**SERPINB2**, **SERPINB8**, **SERPINB10**, **SLPI**	<0.00001 to 0.00081	19 to 49
metalloproteinase inhibitor	**TIMP1**	0.00183	19
decoy receptor	**IL1R2**	0.00077	30
free-heme scavenger molecules	**CD163**, **HP**, **HMOX1**	<0.00001 to 0.00064	26 to 46
complement regulatory molecules	**CD55**, **VSIG4**	<0.00001 to 0.00096	24 to 60
Tissue remodeling and repair genes			
metallopeptidase	**MMP9**	0.00001	33
extracellular matrix components	**COL1A2**[b], **COL8A2**[b], **COL14A1**[b], **COL17A1**[b], **FN1**[b], **SDC1**[b], **VCAN**[b]	<0.00001 to 0.00309	18 to 34
pro-angiogenic factors	**ANG**, **VEGFA**	0.00004 to 0.00236	25 to 30
others	**ARG1**, NOS3	<0.00001 to 0.00054	18 to 44

HUGO gene names are indicated. When genes were represented by several clones on the microarray, p-value and variance medians were calculated. Genes in regular and bold are respectively under- and over-expressed in dengue shock syndrome patients.
[a]percentage of variance associated to disease phenotype.
[b]Danger-associated molecular pattern (DAMP) activity.
doi:10.1371/journal.pone.0011671.t003

Table 4. Pro-inflammatory innate immunity-related genes present in the DSS-gene signature.

Function	Gene Symbol	P-value	Var.[a] (%)	Main cellular origin	Ref
microbicidal peptides	**DEFA1**[b], **DEFA3**[b], **DEFA4**[b]	<0.00001 to 0.00007	0.25 to 0.44	PMN neutro, EpC	[49,50]
	CAMP[b]	<0.00001	0.34	PMN neutro, Mo, mast cells, EpC	[49,50]
	LTF[b]	<0.00001	0.41	PMN neutro, inflammed EpC	[50]
calgranulin proteins	**S100A8**[b], **S100A9**[b]	<0.00001 to 0.00014	0.18 to 0.38	PMN neutro, Mo/Mac	[51,52]
	S100A12[b]	<0.00001	0.33	PMN neutro	[51]
granulocyte enzymes	**RNASE2**[b]	0.00017	0.25	Mo/Mac, Eo, EpC, PMN neutro	[53]
	MPO[b]	0.00024	0.25	PMN neutro, Mo, subtypes of tissue Mac	[50]
	RNASE3[b]	<0.00001	0.29	Eo, Mo, PMN neutro	[54]
	MMP8	<0.00001	0.49	PMN neutro	[50]
	CTSG	<0.00001	0.36	PMN neutro	[50]
	ELANE	<0.00001	0.39	PMN neutro	[50]
pro-inflammatory cytokines and related molecules	**IL18**	0.00052	0.21	Kupffer cells, activated Mac, Mo, DC, EpC	[55]
	IL18BP	0.00710	0.20	T cells, peripheral blood leukocytes, EC	[55]

HUGO gene names are indicated. When genes were represented by several clones on the microarray, p-value and variance medians were calculated. Genes in regular and bold are respectively under- and over-expressed in dengue shock syndrome patients. DC, dendritic cell; EC, endothelial cell; Eo, eosinophil; EpC, epithelial cell; Mac, macrophage, Mo, monocyte; PMN neutro, polymorphonuclear neutrophil; RAGE, receptor for advanced glycation end products.
[a]percentage of variance associated to disease phenotype.
[b]Danger-associated molecular pattern (DAMP) activity.
doi:10.1371/journal.pone.0011671.t004

PLA2G4A, which is the initial rate-limiting enzyme that cleaves membrane phospholipids [78], is over-expressed. Similarly, most downstream key enzymes from the COX-2 and 5-LOX sub-pathways involved in the final synthesis, conversion and transport of inflammatory eicosanoids lipid mediators, are over-expressed. In particular, the transcript encoding the inducible microsomal prostaglandin E synthase PTGES that catalyzes the conversion of prostaglandin PGH2 to PGE2 in the COX-2 sub-pathway, and thought to play a pathogenic role in a number of inflammatory processes [88] is significantly increased and has the highest statistical association with the disease phenotype (62% of gene variance explained by the disease phenotype according to multi-way ANOVA). At the opposite the *PTGDS* transcript, which encodes the anti-inflammatory prostaglandin D2 synthase, has decreased abundance, a finding already reported in metabolic inflammatory processes [89]. Increased abundance of the transcript encoding the LTA4H enzyme that converts the LTA$_4$ leukotriene to LTB$_4$ reflects the activation of the 5-LOX sub-pathway. Finally, transcripts encoding the oxidative enzymes ALOX15B lipoxygenase [84] and cytochrome P450 epoxygenase family members [85], involved in the arachidonic acid metabolic pathway, are also significantly increased in the DSS-gene signature, also reflecting activation of those sub-pathways during DSS.

Thus, a transcriptional signature related to the lipid-related metabolic arachidonic acid pathway is activated in the whole blood cells of DSS children at the time of cardiovascular decompensation.

Discussion

Numerous studies have addressed the pathophysiology of DSS, the more frequent and severe complication of dengue infections. Despite important findings, only partial understanding of the cellular and molecular processes that may support this life-threatening syndrome has been obtained, and we still lack a comprehensive overview of the complete figure of alterations that contribute to – or reflect – the setting-up of the shock syndrome. This could allow the improvement of patients' management and treatment, a major challenge for clinicians. We designed a study aimed at analysing the quasi-global transcriptome of whole blood cells from dengue paediatric patients, looking at every modification that could make sense to the understanding of the pathogenic process. The capacity of such an exhaustive approach to identify relevant host responses, of which unsuspected pathways has been demonstrated in other systemic inflammatory syndromes such as human sepsis or post-trauma sterile SIRS [23–25]. We compared the transcriptome of blood cells from DSS paediatric patients at time of shock to those of children classified as DF or DHF grades I/II [29] matched for important variables such as age, gender, immune status towards dengue infection (primary or secondary infection) and time of disease evolution after onset of fever. Our study has produced significant results, further discussed in the context of DSS pathophysiology.

First, we identify a transcriptional signature of the DSS, differentiating DSS from the other forms of dengue infection and characterizing DSS as a unique and specific entity. Giving particular attention to study design and statistical analysis, we identify a large and robust gene expression profile of 2959 genes that discriminates DSS paediatric patients from other dengue patients, DF or DHF, who did not progress to shock, whatever the supportive treatment they received. Importantly, DSS children clustered together whatever they were considered as having primary or secondary dengue infection, while secondary infections represented the majority of DF, DHF and DSS children recruited (see table S1), as expected in hyper-endemic areas. The robustness of the DSS-associated gene signature was established by showing that the disease phenotype variable significantly affected expression levels of all the genes identified (multi-way ANOVA) and

Table 5. Pro-inflammatory lipid-related genes present in the DSS-gene signature.

Function	Gene Symbol	P-value	Var.[a] (%)	Disease	Ref
Lipid-laden Mo/Mac-related genes					
scavenger receptors of modified LDL in Mo/Mac	**OLR1**, **CD36**, **MSR1**	<0.00001 to 0.00013	0.21 to 0.32	metabolic diseases	[60–64]
lipid nuclear receptor/signalisation by lipids	**PPARG**, PPARA	0.00007 to 0.00732	0.21 to 0.34	metabolic diseases	[65]
efflux of modified cholesterol from Mo/Mac	NPC1	0.00005	0.32	Niemann-Pick disease, atherosclerosis	[66,67]
	ABCA10	0.00016	0.14	none	
migrating Mo/resident Mac chemokine receptors	**CCR2**, CX3CR1	0.00001 to 0.00099	0.22 to 0.40	atherosclerosis	[57]
other lipid-laden-related Mo/Mac genes	**FABP4**, **SOCS6**, **RETN**, **IRS2**	<0.00001 to 0.00092	0.20 to 0.26	metabolic diseases	[68–72]
	CHIT1	<0.00001	0.48	Gaucher's disease, atherosclerosis	[73]
	PCSK9	0.00001	0.42	familial hypercholesterolemia	[74]
	SPP1	<0.00001	0.49	metabolic and inflammatory diseases	[75,76]
anti-oxydant enzymes	LCAT, PAFAH2	0.00196 to 0.00461	21 to 26	metabolic diseases	[77]
Arachidonic acid pathway-related genes					
phospholipase	**PLA2G4A**	0.00003	0.21	rheumatoid arthritis	[78]
eicosanoid synthesis enzymes	**PTGES**, **LTA4H**, PTGDS, **TBXAS1**, PTGDR	<0.00001 to 0.00123	0.22 to 0.63	metabolic and inflammatory diseases, asthma, cancer	[79–82]
leukotrienes convertion enzyme	**MGST2**	0.00003	0.32	none	
leukotriene transporter	**SLCO2B1**	0.00010	0.31	asthma	[83]
lipid oxidation	**ALOX15B**	0.00011	0.33	atherosclerosis	[84]
cytochrome P450 superfamily enzymes	**CYP1B1**, CYP2U1, **CYP51A1**	<0.00001 to 0.00686	10 to 32	Vascular inflammation	[85]

HUGO gene names are indicated. When genes were represented by several clones on the microarray, p-value and variance medians were calculated. Genes in regular and bold are respectively under- and over-expressed in dengue shock syndrome patients.
[a]percentage of variance associated to disease phenotype.
doi:10.1371/journal.pone.0011671.t005

demonstrating the classifying capability of this gene signature using unsupervised hierarchical clustering and SVM/leave-one-out methods [35,39,40]. Based on unsupervised hierarchical clustering, DHF grades I/II patients' expression profiles appear very close or indistinguishable from those of DF patients at the same time of disease evolution, while they group into two heterogeneous sub-groups (1 and 2b, Figure 1), of which significance should be investigated. Altogether, the present results highlight the inadequacy of the 1997 WHO classification of dengue clinical forms [19–21], that considers DF and DHF grades I/II as two separate disease phenotypes, and support the recently proposed classification [90].

Two important questions arise about the DSS-associated transcriptional profile: are the observed modifications of genes expression the cause or the consequence of the pathology, and could these modifications have a predictive value? We cannot definitively answer these questions from the present study, in part because blood samples were collected at the onset of shock (14 out of the 19 DSS patients) or after (5 patients). Functional study of each individual pathway will be required to fully understand the role of each gene in a complex network of molecular interactions.

The ability of some genes transcripts or genes products to accurately predict progression to DSS should be evaluated by multivariate regression models [91] using blood samples collected before the onset of shock, while this proves to be difficult in the context of dengue outbreaks [28]. In the present study, we chose to focus on those of the identified molecular mechanisms that made the more sense to DSS pathophysiology and systemic vascular dysfunction, referring to recent findings on the role of innate immunity in systemic inflammatory processes leading to shock, multi-organ dysfunction syndromes or other pejorative clinical outcomes.

Second, while present results confirm some putative DSS-related biomarkers, it also reveals unreported alterations that make sense to hypovolemic shock pathophysiology. This reinforces the ability of a global and "open mind" approach to identify molecular processes relevant to the studied pathology. Blood cells transcriptional profiles clearly reveal alterations of different immune responses and the activation of a large pro-inflammatory response. A significant proportion of genes of which expression is modified are related to host innate immunity, lymphocyte functions and lipid metabolism in particular. This genome-wide expression analysis also confirms the over-expression of individual biomarkers previously associated with severe dengue, such as the acute phase pentraxin-related protein PTX3, the pro-inflammatory IL-18 cytokine or the anti-inflammatory IL-10 cytokine (Table S2) [11,12,15], providing a more comprehensive overview of their implication in the pathophysiology of DSS.

Our results differ however from those reported by Long *et al* in a genome-wide expression profiling study comparing DSS children with uncomplicated paediatric patients [28]. This study concluded on a global "benign" and "muted" immune transcriptional response but a decreased expression of genes involved in IL-10 and IFN type I-related pathways in DSS children blood cells [28].

Differences in study design, size of cohorts and time of blood sampling from patients in the course of dengue disease may explain these differences. Indeed, in our study, two DSS children had gene expression profiles close to those of uncomplicated DF and DHF, and clustered within the DF/DHF cluster. Both proved to be the children from whom blood was sampled three days after the onset of shock, while the three DSS children sampled two days after shock onset still exhibited a typical DSS gene expression profile. This suggests that a shift from a "severe" to an "uncomplicated" transcriptional profile may occur within a very short time, and could explain the "uncomplicated" and benign gene immune transcriptional responses reported by Long *et al* [28]. Differences in strategies and methods used to filter genes differentially expressed between patients' groups could also explain the finding that few IFN type I-related genes but a large diversity of other pathways were identified in the present study compared to other transcriptomic studies of DHF or DSS patients. Here, genes were selected considering only their statistical significance and their association with the disease phenotype. Differently from others [27,28,42], no "fold change" cut-off filter was applied since this non-statistically-motivated criteria selects preferentially genes prompt to high variations such as the IFN type I-induced genes [92,93], thus excluding from subsequent bio-informatic analysis a diversity of transcripts exhibiting more subtle variations but strong associations and biological relevance with the considered disease phenotype.

Third, unsuspected mechanisms identified in DSS patients could contribute importantly to the pathophysiology of this severe syndrome, as supported by similarities between those DSS-related alterations and other critical syndromes. Interestingly, a number of immune, repair-remodeling and metabolic-related related pathways are simultaneously altered in the blood cells of DSS children at the onset of shock. In particular, T and NK lymphocyte transcriptional responses are globally impaired while genes implicated in compensatory anti-inflammatory and repair/remodeling responses and in innate immune responses are over-expressed. This highlights the complexity of biological responses at the time of dengue shock syndrome, and points out similarities between DSS and other critical syndromes such as severe sepsis, or post-trauma SIRS that are similarly characterized by depressed T lymphocyte responses but over-expressed innate immunity [94,95].

Reduced abundance of a number of T-lymphocyte related transcripts at the time of DSS may reflect a feed-back mechanism aimed at limiting an initial early T lymphocyte activation, reported to occur in patients who further progress to severe dengue [96,97]. Such a negative feed-back may be sustained by the over-expression of a diversity of anti-inflammatory transcripts in DSS patients' blood cells at the time of shock. In particular, the two potent immunomodulating factors prostaglandin E synthase and VSIG4, which dampen both T and NK lymphocyte responses [43–45] and have both a strong statistical association with the DSS phenotype, could have such a negative effect. Based on those observations and previous clinical reports, the benefit of corticotherapy in DSS patients might thus be questioned [98,99].

Over-expression in the blood of DSS children of several repair and remodeling genes encoding extracellular matrix proteins, vasoactive mediators and matrix metalloproteases such as the MMP9, likely reflects a compensatory response to inflammatory insults, and a number of those genes products are now considered putative biomarkers in systemic inflammatory syndromes such as severe sepsis [100]. Most proteins encoded by those genes are indeed secreted by activated immune cells such as monocytes/macrophages. They may have adverse effects towards the vascular endothelium when produced in excess, since they may increase immobilization of inflammatory mediators at the surface of endothelial cells [101], permeability of capillaries [102], or induce direct damage to endothelial tissues [103]. Recently, one of them, MMP9 has been proposed as a putative candidate in the occurrence of plasma leakage during dengue infection [47].

While previous transcriptional studies failed to identify pro-inflammatory gene patterns in the blood cells of DSS patients [27,28,42], our study is the first one to report that a diversity of pro-inflammatory transcriptional responses at the interface of innate immunity, inflammation and host lipid metabolism are activated at the time of cardiovascular failure. Since those mechanisms are considered pathogenic in other systemic inflammatory diseases where systemic vascular dysfunction does occur, we suggest that they may altogether contribute to DSS pathophysiology.

Activation of a pro-inflammatory defence gene pattern in DSS patients' blood cells (Table 4) has relevance to the pathophysiology of systemic vascular dysfunction, since most microbicidal peptides and enzymes have recognized pro-inflammatory and pathogenic effects towards vascular endothelial tissues [50]. Among them, the neutrophil microbicidal peptides alpha defensins and the highly pro-inflammatory calgranulins proteins S100A8/A9 and S100A12 are now considered putative pathogenic factors in sepsis, cardiovascular diseases, rheumatoid arthritis or atherosclerosis [51,56,104]. While neutrophils are considered the main source of those defence molecules, this cellular origin cannot be established from the present study due to the cellular complexity of unfractionnated whole blood samples and to the possibility that other circulating cell types may express a neutrophil-like inflammatory repertoire under pathologic conditions [105]. A putative neutrophil origin of this gene expression pattern is however supported by the over-expression in DSS patients blood cells of transcripts encoding other neutrophil-related molecules such as the MMP8 matrix metalloprotease, CEACAM-6, CEACAM-8 and CD99L2 adhesion molecules (Tables 4 and S2) involved in the recruitment of neutrophils to vascular endothelia. Functional studies should confirm whether those first line-defence immune cells which produce an array of pro-inflammatory mediators highly damaging to host tissues and vascular endothelia [50,106], and poorly regarded in dengue [107], definitively play a role in DSS pathophysiology.

Alteration of a gene pattern related to homeostasis of cholesterol in monocytes/macrophages (Mo/Mac) in the blood cells of DSS children (Table 5) was an unexpected finding, while it should be considered regarding recent knowledge on the role of monocytes as a pivotal link between inflammation, innate immunity and host lipid metabolism [108,109]. Indeed, under physiological conditions, monocytes maintain cholesterol homeostasis by clearing modified LDL such as oxidized LDL (ox-LDL) from plasma. Under pathological conditions, balance between uptake and efflux of those modified cholesterol molecules may be altered [59,60,62,66,67,86], resulting in the intracellular accumulation of modified cholesterol. This turns classical monocytes towards a pro-inflammatory phenotype, lipid-laden monocytes/macrophages (Mo/Mac), a sub-type of pro-inflammatory immune cells initially identified in vascular lesions of chronic inflammatory metabolic diseases [110].

Recent knowledge has shown that those atypical monocytes produce a large array of pro-inflammatory mediators such as ROS, metalloproteases, eicosanoids, and pro-inflammatory adipokines, making these cells potent contributors to vascular damages, systemic inflammation and major metabolic changes

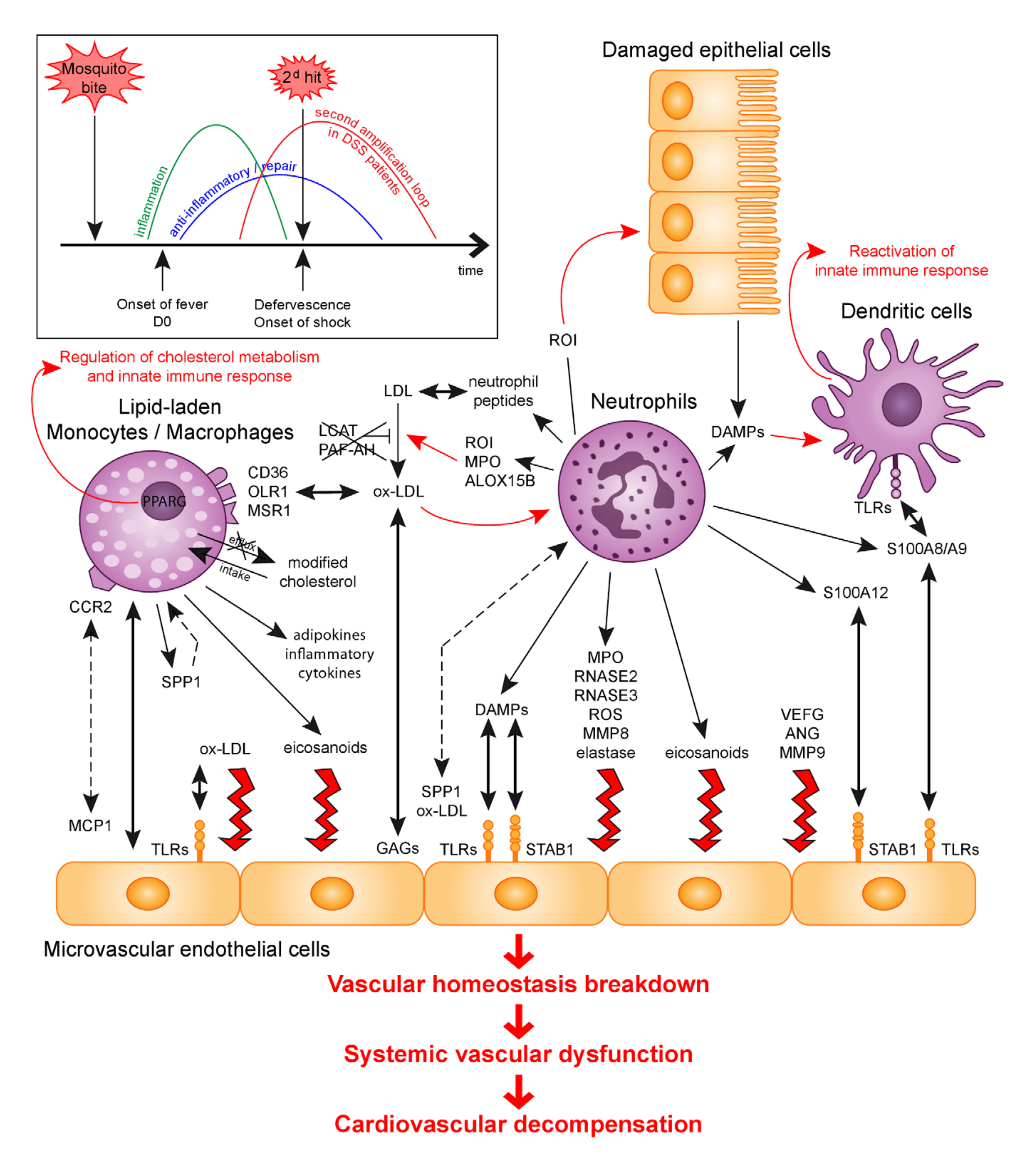


Figure 4. Hypothesis of a second inflammatory amplification loop in dengue shock syndrome. After induction of a first inflammatory and anti-viral response to dengue virus, disease resolution generally occurs around time of defervescence for most dengue-infected patients. Some patients however progress towards a life-threatening dengue shock syndrome. Results obtained in this study suggest that in those patients, a second inflammatory amplification loop, which involves a diversity of pro-inflammatory responses related to innate immunity, occurs and leads to a major inflammatory systemic syndrome and to vascular homeostasis breakdown. The putative role of different markers identified in vascular endothelial dysfunction is indicated. Thin black arrow, release of; Bold black arrow, interaction between; Punctuated black arrow, chemotactic effect; Thin red arrow, biological activity; Bold red arrow, direct activity on endothelium. DAMPs, danger-associated molecular pattern; GAG, glycosaminoglycane; ROI, reactive oxygen intermediates; TLR, Toll-like receptor.
doi:10.1371/journal.pone.0011671.g004

such as insulin-resistance [109,111], which altogether characterize systemic inflammatory syndromes such as DSS or severe sepsis. While the existence of functional lipid-laden Mo/Mac during DSS should be established by functional studies, such a molecular mechanism could explain decrease of circulating sub-fractions and total cholesterol previously reported in DSS [112–115] and in other critically ill patients where low cholesterol levels are associated with poor clinical outcome [116,117].

Altered homeostasis of cholesterol in blood cells from DSS patients could also favour replication of dengue viruses into host cells [118], thus contributing to increased viremia in patients with severe dengue infection [119,120], while this could not be evaluated in this study since part of patients had undetectable viremia at the time of blood sampling.

The factors contributing to altered homeostasis of cholesterol in the blood cells of DSS children at time of shock are numerous. Increased lipid peroxydation activity [113,114] insufficiently compensated by anti-oxydant mechanisms [121], as supported by related altered gene patterns identified in this study (Table 5), may result in high levels of circulating ox-LDL contributing to altered cholesterol metabolism. Differences in nutritional status [122–124] or host genetics may also contribute to altered homeostasis of cholesterol gene pattern in the blood cells of DSS patients. Interestingly, transcripts encoding molecules considered candidates to diseases characterized by impairment of cholesterol homeostasis such as *NPC1*, *PCSK9* and *PPARG* [66,67,74], have significant altered abundance in the blood cells of DSS children (Table 5). Further investigations should consider possible associations between DSS and allelic variants of such genes. Whatever the determinants of cholesterol metabolism alterations in DSS patients, our results reinforce interest in considering sub-fractions and total cholesterol as putative biomarkers of DSS [115]. They also suggest that drugs used to treat metabolic disorders such as atherosclerosis should deserve further attention for the control of such a pro-inflammatory process in dengue-infected patients, now proposed for other critical illnesses [117].

Transcriptional activation of the lipid-related arachidonic acid pathway in the whole blood cells of DSS children at the time of shock was another pro-inflammatory mechanism relevant to the pathophysiology of DSS [81]. Activation of this lipid metabolic pathway in innate cells such as neutrophils or lipid-laden monocytes during inflammatory process or infection [125], results in the production of eicosanoid lipid mediators, that are not only physiological regulators of vascular tone and permeability [81] but also potent pro-inflammatory mediators involved in a number of pathologies such as asthma [81]. Interestingly, formation of lipid bodies where eicosanoid synthesis takes place, can be induced by ox-LDL through activation of the PPARγ nuclear-lipid receptor [126], thus suggesting a direct link between the three pro-inflammatory pathways identified in DSS children and a contribution of arachidonic pathway-related inflammatory lipids and oxidative enzymes to the systemic vascular dysfunction leading to DSS.

Fourth, DAMPs and TLRs could be a link from primary to secondary inflammation, leading to DSS. Occurrence of DSS in only some patients at the late phase of infection is likely due to an inadequate control or an amplification of the primary inflammatory response aimed at fighting infection. The pro-inflammatory molecular responses activated in the blood cells of DSS children at time of shock involve a diversity of innate immune mediators that may amplify a first-line inflammatory response mediated by TNF, IL-6 or IL-1, thus contributing to a secondary inflammatory loop. Indeed, a number of repair/remodeling and of defence gene products over-expressed in DSS patients blood cells are considered endogenous danger signals or Danger-Associated Molecular Patterns (DAMPs) (Tables 3 and 4; molecules with DAMP activity are indicated) capable to trigger secondary systemic inflammatory responses through direct interaction with surface or intracellular receptors such as TLRs or NODs expressed in endothelial or innate immune cells [127]. DAMPs include a diversity of molecules without structural similarity either actively produced by immune cells in the context of an infection or passively secreted by damaged tissues [128,129], now considered key inducers of secondary systemic inflammation in a number of acute inflammatory syndromes [130,131] or chronic diseases [132].

Amplification of inflammation during DSS through direct signalling by molecules harbouring DAMP activity via TLRs, is also supported by the increased abundance of DAMP-induced transcripts as those encoding the pro-inflammatory IL-18 cytokine or the NLRC4/CARD12 intracellular sensor [55]. Interestingly, association of allelic polymorphisms of TLR4 with DSS suggested by De Kruif and colleagues [46] suggests that differential signalling through TLRs may contribute to the severity of dengue disease, as suspected for other pathologies [133]. Accordingly, anti-inflammatory drugs targeting Toll-like receptors are now under development for a number of inflammatory pathologies where innate immunity and TLRs play a central pathogenic role [134].

DSS pathophysiology: a secondary inflammatory loop hypothesis

To summarize, we report the identification of a specific gene expression profile in the blood cells of DSS children at time of shock, characterizing DSS as a unique entity at the transcriptional level whatever the immunological status of children regarding primary or secondary infection. Major immunological alterations identified at the time of shock are characterized by an altered balance between depressed T lymphocyte responses and exacerbated compensatory and pro-inflammatory innate immune responses that may, finally, be detrimental to the host [135–137], while functional studies should confirm the contribution of those molecular mechanisms to DSS pathophysiology.

Based on recent knowledge on molecular mechanisms altered in other systemic inflammatory diseases, DSS may result from a complex pro-inflammatory network involving a diversity of innate immune effectors sustaining a secondary systemic inflammatory loop, leading in turn to vascular homeostasis breakdown and systemic microcirculatory failure characterizing DSS (Figure 4).

We suggest that drugs available to treat metabolic and other systemic chronic inflammatory diseases could be considered for the treatment of dengue-infected patients before shock occurs, and that a number of bio-markers found altered in DSS patients blood cells should be evaluated. as putative predictive markers of progression to DSS.

Supporting Information

Figure S1 Validation of microarray results by RT-PCR. Pearson's correlation was calculated between microarray expression signals (horizontal axis) and Delta Ct values from real-time PCR (vertical axis) for nine genes highly associated to dengue shock syndrome. ** Correlation is significant at 0.01.
Found at: doi:10.1371/journal.pone.0011671.s001 (4.94 MB TIF)

Table S1 Clinical and biological characteristics of each DF, DHF and DSS patient
Found at: doi:10.1371/journal.pone.0011671.s002 (0.04 MB XLS)

Table S2 List of the 3515 clones corresponding to the 2959 genes differentially expressed between DF, DHF and DSS

patients, identified using the multi-way ANOVA at a false discovery rate of 10. Clones corresponding to the 2959 genes are listed according to their association to DSS, the first one being the gene of which expression level variance is the most influenced by the clinical phenotype. HUGO gene names are indicated. The variation is the one related to the DSS group relatively to DF and DHF. ANOVA, analysis of variance; DF, dengue fever; DHF, dengue hemorrhagic fever; DSS, dengue shock syndrome; NA, not available. a percentage of variance associated to disease phenotype.
Found at: doi:10.1371/journal.pone.0011671.s003 (0.90 MB XLS)

Acknowledgments

We greatly thank Pr Y. Buisson for supporting this program, I. Drouet and H. Puggelli for technical support and help in preparation of field work.

We thank Dr J. Desplans for help in using the Agilent microarrays platform.

We are also indebted to all doctors, nurses, patients and their families who participated to this study at the hospital of Kampong Cham, Cambodia, and to all technicians from the virology department of the Institut Pasteur in Cambodia who carried out dengue diagnosis assays.

Author Contributions

Conceived and designed the experiments: SD CS VD PB PCP. Performed the experiments: SD CS PCP. Analyzed the data: SD CS AB PR PCP. Contributed reagents/materials/analysis tools: VD SO PTL NC SN PB PCP. Wrote the paper: SD PCP. Contributed to obtainment of funding: HT.

References

1. Gubler DJ (2002) Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. Trends Microbiol 10: 100–103.
2. Khun S, Manderson L (2008) Poverty, user fees and ability to pay for health care for children with suspected dengue in rural Cambodia. Int J Equity Health 7: 10.
3. Peters KG (1998) Vascular endothelial growth factor and the angiopoietins: working together to build a better blood vessel. Circ Res 83: 342–343.
4. Basu A, Chaturvedi UC (2008) Vascular endothelium: the battlefield of dengue viruses. FEMS Immunol Med Microbiol 53: 287–299.
5. Pang T, Cardosa MJ, Guzman MG (2007) Of cascades and perfect storms: the immunopathogenesis of dengue haemorrhagic fever-dengue shock syndrome (DHF/DSS). Immunol Cell Biol 85: 43–45.
6. Green S, Rothman A (2006) Immunopathological mechanisms in dengue and dengue hemorrhagic fever. Curr Opin Infect Dis 19: 429–436.
7. Lin CF, Wan SW, Cheng HJ, Lei HY, Lin YS (2006) Autoimmune pathogenesis in dengue virus infection. Viral Immunol 19: 127–132.
8. Murgue B (2009) Severe dengue: questioning the paradigm. Microbes Infect 12: 113–118.
9. Fink J, Gu F, Vasudevan SG (2006) Role of T cells, cytokines and antibody in dengue fever and dengue haemorrhagic fever. Rev Med Virol 16: 263–275.
10. Libraty DH, Acosta LP, Tallo V, Segubre-Mercado E, Bautista A, et al. (2009) A prospective nested case-control study of Dengue in infants: rethinking and refining the antibody-dependent enhancement dengue hemorrhagic fever model. PLoS Med 6: e1000171.
11. Green S, Vaughn DW, Kalayanarooj S, Nimmannitya S, Suntayakorn S, et al. (1999) Elevated plasma interleukin-10 levels in acute dengue correlate with disease severity. J Med Virol 59: 329–334.
12. Mustafa AS, Elbishbishi EA, Agarwal R, Chaturvedi UC (2001) Elevated levels of interleukin-13 and IL-18 in patients with dengue hemorrhagic fever. FEMS Immunol Med Microbiol 30: 229–233.
13. Juffrie M, Meer GM, Hack CE, Haasnoot K, Sutaryo, et al. (2001) Inflammatory mediators in dengue virus infection in children: interleukin-6 and its relation to C-reactive protein and secretory phospholipase A2. Am J Trop Med Hyg 65: 70–75.
14. Koraka P, Murgue B, Deparis X, Van Gorp EC, Setiati TE, et al. (2004) Elevation of soluble VCAM-1 plasma levels in children with acute dengue virus infection of varying severity. J Med Virol 72: 445–450.
15. Mairuhu AT, Peri G, Setiati TE, Hack CE, Koraka P, et al. (2005) Elevated plasma levels of the long pentraxin, pentraxin 3, in severe dengue virus infections. J Med Virol 76: 547–552.
16. Cardier JE, Marino E, Romano E, Taylor P, Liprandi F, et al. (2005) Proinflammatory factors present in sera from patients with acute dengue infection induce activation and apoptosis of human microvascular endothelial cells: possible role of TNF-alpha in endothelial cell damage in dengue. Cytokine 30: 359–365.
17. Tseng CS, Lo HW, Teng HC, Lo WC, Ker CG (2005) Elevated levels of plasma VEGF in patients with dengue hemorrhagic fever. FEMS Immunol Med Microbiol 43: 99–102.
18. Bozza FA, Cruz OG, Zagne SM, Azeredo EL, Nogueira RM, et al. (2008) Multiplex cytokine profile from dengue patients: MIP-1beta and IFN-gamma as predictive factors for severity. BMC Infect Dis 8: 86.
19. Deen JL, Harris E, Wills B, Balmaseda A, Hammond SN, et al. (2006) The WHO dengue classification and case definitions: time for a reassessment. Lancet 368: 170–173.
20. Bandyopadhyay S, Lum LC, Kroeger A (2006) Classifying dengue: a review of the difficulties in using the WHO case classification for dengue haemorrhagic fever. Trop Med Int Health 11: 1238–1255.
21. Ng CF, Lum LC, Ismail NA, Tan LH, Tan CP (2007) Clinicians' diagnostic practice of dengue infections. J Clin Virol 40: 202–206.
22. Cobb JP, Mindrinos MN, Miller-Graziano C, Calvano SE, Baker HV, et al. (2005) Application of genome-wide expression analysis to human health and disease. Proc Natl Acad Sci U S A 102: 4801–4806.
23. Feezor RJ, Cheng A, Paddock HN, Baker HV, Moldawer LL (2005) Functional genomics and gene expression profiling in sepsis: beyond class prediction. Clin Infect Dis 41 Suppl 7: S427–435.
24. Tang BM, McLean AS, Dawes IW, Huang SJ, Lin RC (2007) The use of gene-expression profiling to identify candidate genes in human sepsis. Am J Respir Crit Care Med 176: 676–684.
25. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, et al. (2005) Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci U S A 102: 13544–13549.
26. Laudanski K, Miller-Graziano C, Xiao W, Mindrinos MN, Richards DR, et al. (2006) Cell-specific expression and pathway analyses reveal alterations in trauma-related human T cell and monocyte pathways. Proc Natl Acad Sci U S A 103: 15564–15569.
27. Simmons CP, Popper S, Dolocek C, Chau TN, Griffiths M, et al. (2007) Patterns of host genome-wide gene transcript abundance in the peripheral blood of patients with acute dengue hemorrhagic fever. J Infect Dis 195: 1097–1107.
28. Long HT, Hibberd ML, Hien TT, Dung NM, Van Ngoc T, et al. (2009) Patterns of gene transcript abundance in the blood of children with severe or uncomplicated dengue highlight differences in disease evolution and host response to dengue virus infection. J Infect Dis 199: 537–546.
29. World Health Organization (1997) Dengue haemorrhagic fever: diagnosis, treatment, prevention and control. Geneva: World Health Organization. 84 p.
30. Buchy P, Vo VL, Bui KT, Trinh TX, Glaziou P, et al. (2005) Secondary dengue virus type 4 infections in Vietnam. Southeast Asian J Trop Med Public Health 36: 178–185.
31. Reynes JM, Ong S, Mey C, Ngan C, Hoyer S, et al. (2003) Improved molecular detection of dengue virus serotype 1 variants. J Clin Microbiol 41: 3864–3867.
32. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207–210.
33. Smyth GK, Michaud J, Scott HS (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics 21: 2067–2075.
34. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185–193.
35. Didier G, Brezellec P, Remy E, Henaut A (2002) GeneANOVA–gene expression analysis of variance. Bioinformatics 18: 490–491.
36. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: a pratical and powerful approach to multiple testing. J Royal Stat Soc Ser B 57: 289–300.
37. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863–14868.
38. Saldanha AJ (2004) Java Treeview–extensible visualization of microarray data. Bioinformatics 20: 3246–3248.
39. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, et al. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A 97: 262–267.
40. Kerr MK, Martin M, Churchill GA (2000) Analysis of variance for gene expression microarray data. J Comput Biol 7: 819–837.

41. Pavlidis P (2003) Using ANOVA for gene selection from microarray studies of the nervous system. Methods 31: 282–289.
42. Ubol S, Masrinoul P, Chaijaruwanich J, Kalayanarooj S, Charoensirisuthikul T, et al. (2008) Differences in global gene expression in peripheral blood mononuclear cells indicate a significant role of the innate responses in progression of dengue fever but not dengue hemorrhagic fever. J Infect Dis 197: 1459–1467.
43. Stobo JD, Kennedy MS, Goldyne ME (1979) Prostaglandin E modulation of the mitogenic response of human T cells. Differential response of T-cell subpopulations. J Clin Invest 64: 1188–1203.
44. Joshi PC, Zhou X, Cuchens M, Jones Q (2001) Prostaglandin E2 suppressed IL-15-mediated human NK cell function through down-regulation of common gamma-chain. J Immunol 166: 885–891.
45. Vogt L, Schmitz N, Kurrer MO, Bauer M, Hinton HI, et al. (2006) VSIG4, a B7 family-related protein, is a negative regulator of T cell activation. J Clin Invest 116: 2817–2826.
46. de Kruif MD, Setiati TE, Mairuhu AT, Koraka P, Aberson HA, et al. (2008) Differential gene expression changes in children with severe dengue virus infections. PLoS Negl Trop Dis 2: e215.
47. Luplertlop N, Misse D, Bray D, Deleuze V, Gonzalez JP, et al. (2006) Dengue-virus-infected dendritic cells trigger vascular leakage through metalloproteinase overproduction. EMBO Rep 7: 1176–1181.
48. Predescu D, Predescu S, Shimizu J, Miyawaki-Shimizu K, Malik AB (2005) Constitutive eNOS-derived nitric oxide is a determinant of endothelial junctional integrity. Am J Physiol Lung Cell Mol Physiol 289: L371–381.
49. Yang D, Biragyn A, Hoover DM, Lubkowski J, Oppenheim JJ (2004) Multiple roles of antimicrobial defensins, cathelicidins, and eosinophil-derived neurotoxin in host defense. Annu Rev Immunol 22: 181–215.
50. DiStasi MR, Ley K (2009) Opening the flood-gates: how neutrophil-endothelial interactions regulate permeability. Trends Immunol 30: 547–556.
51. Foell D, Wittkowski H, Vogl T, Roth J (2007) S100 proteins expressed in phagocytes: a novel group of damage-associated molecular pattern molecules. J Leukoc Biol 81: 28–37.
52. Vogl T, Tenbrock K, Ludwig S, Leukert N, Ehrhardt C, et al. (2007) Mrp8 and Mrp14 are endogenous activators of Toll-like receptor 4, promoting lethal, endotoxin-induced shock. Nat Med 13: 1042–1049.
53. Yang D, Chen Q, Rosenberg HF, Rybak SM, Newton DL, et al. (2004) Human ribonuclease A superfamily members, eosinophil-derived neurotoxin and pancreatic ribonuclease, induce dendritic cell maturation and activation. J Immunol 173: 6134–6142.
54. Sur S, Glitz DG, Kita H, Kujawa SM, Peterson EA, et al. (1998) Localization of eosinophil-derived neurotoxin and eosinophil cationic protein in neutrophilic leukocytes. J Leukoc Biol 63: 715–722.
55. Pedra JH, Cassel SL, Sutterwala FS (2009) Sensing pathogens and danger signals by the inflammasome. Curr Opin Immunol 21: 10–16.
56. Foell D, Frosch M, Sorg C, Roth J (2004) Phagocyte-specific calcium-binding S100 proteins as clinical laboratory markers of inflammation. Clin Chim Acta 344: 37–51.
57. Gordon S (2007) Macrophage heterogeneity and tissue lipids. J Clin Invest 117: 89–93.
58. Barlic J, Murphy PM (2007) Chemokine regulation of atherosclerosis. J Leukoc Biol 82: 226–236.
59. Mosig S, Rennert K, Buttner P, Krause S, Lutjohann D, et al. (2008) Monocytes of patients with familial hypercholesterolemia show alterations in cholesterol metabolism. BMC Med Genomics 1: 60.
60. de Winther MP, van Dijk KW, Havekes LM, Hofker MH (2000) Macrophage scavenger receptor class A: A multifunctional receptor in atherosclerosis. Arterioscler Thromb Vasc Biol 20: 290–297.
61. Febbraio M, Hajjar DP, Silverstein RL (2001) CD36: a class B scavenger receptor involved in angiogenesis, atherosclerosis, inflammation, and lipid metabolism. J Clin Invest 108: 785–791.
62. Pennings M, Meurs I, Ye D, Out R, Hoekstra M, et al. (2006) Regulation of cholesterol homeostasis in macrophages and consequences for atherosclerotic lesion development. FEBS Lett 580: 5588–5596.
63. Chen XP, Zhang TT, Du GH (2007) Lectin-like oxidized low-density lipoprotein receptor-1, a new promising target for the therapy of atherosclerosis? Cardiovasc Drug Rev 25: 146–161.
64. Renie G, Maingrette F, Li L (2007) Diabetic vasculopathy and the lectin-like oxidized low-density lipoprotein receptor-1 (LOX-1). Curr Diabetes Rev 3: 103–110.
65. Szanto A, Roszer T (2008) Nuclear receptors in macrophages: a link between metabolism and inflammation. FEBS Lett 582: 106–116.
66. Ory DS (2004) The niemann-pick disease genes; regulators of cellular cholesterol homeostasis. Trends Cardiovasc Med 14: 66–72.
67. Zhang JR, Coleman T, Langmade SJ, Scherrer DE, Lane L, et al. (2008) Niemann-Pick C1 protects against atherosclerosis in mice via regulation of macrophage intracellular cholesterol trafficking. J Clin Invest 118: 2281–2290.
68. Sesti G, Federici M, Hribal ML, Lauro D, Sbraccia P, et al. (2001) Defects of the insulin receptor substrate (IRS) system in human metabolic disorders. Faseb J 15: 2099–2111.
69. White MF (2002) IRS proteins and the common path to diabetes. Am J Physiol Endocrinol Metab 283: E413–422.
70. Makowski L, Hotamisligil GS (2005) The role of fatty acid binding proteins in metabolic syndrome and atherosclerosis. Curr Opin Lipidol 16: 543–548.
71. Howard JK, Flier JS (2006) Attenuation of leptin and insulin signaling by SOCS proteins. Trends Endocrinol Metab 17: 365–371.
72. Qatanani M, Szwergold NR, Greaves DR, Ahima RS, Lazar MA (2009) Macrophage-derived human resistin exacerbates adipose tissue inflammation and insulin resistance in mice. J Clin Invest.
73. Boot RG, van Achterberg TA, van Aken BE, Renkema GH, Jacobs MJ, et al. (1999) Strong induction of members of the chitinase family of proteins in atherosclerosis: chitotriosidase and human cartilage gp-39 expressed in lesion macrophages. Arterioscler Thromb Vasc Biol 19: 687–694.
74. Horton JD, Cohen JC, Hobbs HH (2009) PCSK9: a convertase that coordinates LDL catabolism. J Lipid Res 50 Suppl: S172–177.
75. Scatena M, Liaw L, Giachelli CM (2007) Osteopontin: a multifunctional molecule regulating chronic inflammation and vascular disease. Arterioscler Thromb Vasc Biol 27: 2302–2309.
76. Cho HJ, Cho HJ, Kim HS (2009) Osteopontin: a multifunctional protein at the crossroads of inflammation, atherosclerosis, and vascular calcification. Curr Atheroscler Rep 11: 206–213.
77. Forte TM, Subbanagounder G, Berliner JA, Blanche PJ, Clermont AO, et al. (2002) Altered activities of anti-atherogenic enzymes LCAT, paraoxonase, and platelet-activating factor acetylhydrolase in atherosclerosis-susceptible mice. J Lipid Res 43: 477–485.
78. Gilroy DW, Newson J, Sawmynaden P, Willoughby DA, Croxtall JD (2004) A novel role for phospholipase A2 isoforms in the checkpoint control of acute inflammation. Faseb J 18: 489–498.
79. Homaidan FR, Chakroun I, Haidar HA, El-Sabban ME (2002) Protein regulators of eicosanoid synthesis: role in inflammation. Curr Protein Pept Sci 3: 467–484.
80. Khanapure SP, Garvey DS, Janero DR, Letts LG (2007) Eicosanoids in inflammation: biosynthesis, pharmacology, and therapeutic frontiers. Curr Top Med Chem 7: 311–340.
81. Harizi H, Corcuff JB, Gualde N (2008) Arachidonic-acid-derived eicosanoids: roles in biology and immunopathology. Trends Mol Med 14: 461–469.
82. Rossi A, Cuzzocrea S, Sautebin L (2009) Involvement of leukotriene pathway in the pathogenesis of ischemia-reperfusion injury and septic and non-septic shock. Curr Vasc Pharmacol 7: 185–197.
83. Lima JJ, Blake KV, Tantisira KG, Weiss ST (2009) Pharmacogenetics of asthma. Curr Opin Pulm Med 15: 57–62.
84. Wittwer J, Hersberger M (2007) The two faces of the 15-lipoxygenase in atherosclerosis. Prostaglandins Leukot Essent Fatty Acids 77: 67–77.
85. Deng Y, Theken KN, Lee CR (2009) Cytochrome P450 epoxygenases, soluble epoxide hydrolase, and the regulation of cardiovascular inflammation. J Mol Cell Cardiol 48: 331–341.
86. Attie AD, Kastelein JP, Hayden MR (2001) Pivotal role of ABCA1 in reverse cholesterol transport influencing HDL levels and susceptibility to atherosclerosis. J Lipid Res 42: 1717–1726.
87. Hennuyer N, Tailleux A, Torpier G, Mezdour H, Fruchart JC, et al. (2005) PPARalpha, but not PPARgamma, activators decrease macrophage-laden atherosclerotic lesions in a nondiabetic mouse model of mixed dyslipidemia. Arterioscler Thromb Vasc Biol 25: 1897–1902.
88. Samuelsson B, Morgenstern R, Jakobsson PJ (2007) Membrane prostaglandin E synthase-1: a novel therapeutic target. Pharmacol Rev 59: 207–224.
89. Cipollone F, Fazia M, Iezzi A, Ciabattoni G, Pini B, et al. (2004) Balance between PGD synthase and PGE synthase is a major determinant of atherosclerotic plaque instability in humans. Arterioscler Thromb Vasc Biol 24: 1259–1265.
90. World Health Organization (2009) Dengue guidelines for diagnosis, treatment, prevention and control: new edition. Geneva: World Health Organization. 147 p.
91. Dyrskjot L, Zieger K, Real FX, Malats N, Carrato A, et al. (2007) Gene expression signatures predict outcome in non-muscle-invasive bladder carcinoma: a multicenter validation study. Clin Cancer Res 13: 3545–3551.
92. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, et al. (2003) Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci U S A 100: 1896–1901.
93. Eady JJ, Wortley GM, Wormstone YM, Hughes JC, Astley SB, et al. (2005) Variation in gene expression profiles of peripheral blood mononuclear cells from healthy volunteers. Physiol Genomics 22: 402–411.
94. Remick DG (2007) Pathophysiology of sepsis. Am J Pathol 170: 1435–1444.
95. Lenz A, Franklin GA, Cheadle WG (2007) Systemic inflammation after trauma. Injury 38: 1336–1345.
96. Chaturvedi UC, Shrivastava R, Tripathi RK, Nagar R (2007) Dengue virus-specific suppressor T cells: current perspectives. FEMS Immunol Med Microbiol 50: 285–299.
97. Green S, Vaughn DW, Kalayanarooj S, Nimmannitya S, Suntayakorn S, et al. (1999) Early immune activation in acute dengue illness is related to development of plasma leakage and disease severity. J Infect Dis 179: 755–762.
98. Panpanich R, Sornchai P, Kanjanaratanakorn K (2006) Corticosteroids for treating dengue shock syndrome. Cochrane Database Syst Rev 3: CD003488.
99. Rajapakse S (2009) Corticosteroids in the treatment of dengue illness. Trans R Soc Trop Med Hyg 103: 122–126.
100. Lorente L, Martin MM, Sole-Violan J, Blanquer J, Paramo JA (2010) Matrix metalloproteinases and their inhibitors as biomarkers of severity in sepsis. Crit Care 14: 402.



101. Wight TN, Merrilees MJ (2004) Proteoglycans in atherosclerosis and restenosis: key roles for versican. Circ Res 94: 1158–1167.
102. Galley HF, Webster NR (2004) Physiology of the endothelium. Br J Anaesth 93: 105–113.
103. Raffetto JD, Khalil RA (2008) Matrix metalloproteinases and their inhibitors in vascular remodeling and vascular disease. Biochem Pharmacol 75: 346–359.
104. Ehrchen JM, Sunderkotter C, Foell D, Vogl T, Roth J (2009) The endogenous Toll-like receptor 4 agonist S100A8/S100A9 (calprotectin) as innate amplifier of infection, autoimmunity, and cancer. J Leukoc Biol 86: 557–566.
105. Borregaard N, Sorensen OE, Theilgaard-Monch K (2007) Neutrophil granules: a library of innate immunity proteins. Trends Immunol 28: 340–345.
106. Quinn K, Henriques M, Parker T, Slutsky AS, Zhang H (2008) Human neutrophil peptides: a novel potential mediator of inflammatory cardiovascular diseases. Am J Physiol Heart Circ Physiol 295: H1817–1824.
107. Butthep P, Bunyaratvej A, Bhamarapravati N (1993) Dengue virus and endothelial cell: a related phenomenon to thrombocytopenia and granulocytopenia in dengue hemorrhagic fever. Southeast Asian J Trop Med Public Health 24 Suppl 1: 246–249.
108. Michelsen KS, Doherty TM, Shah PK, Arditi M (2004) TLR signaling: an emerging bridge from innate immunity to atherogenesis. J Immunol 173: 5901–5907.
109. Wellen KE, Hotamisligil GS (2005) Inflammation, stress, and diabetes. J Clin Invest 115: 1111–1119.
110. Schmitz G, Grandl M (2008) Lipid homeostasis in macrophages - implications for atherosclerosis. Rev Physiol Biochem Pharmacol 160: 93–125.
111. Yagmur E, Trautwein C, Gressner AM, Tacke F (2006) Resistin serum levels are associated with insulin resistance, disease severity, clinical complications, and prognosis in patients with chronic liver diseases. Am J Gastroenterol 101: 1244–1252.
112. van Gorp EC, Suharti C, Mairuhu AT, Dolmans WM, van Der Ven J, et al. (2002) Changes in the plasma lipid profile as a potential predictor of clinical outcome in dengue hemorrhagic fever. Clin Infect Dis 34: 1150–1153.
113. Lee CY, Seet RC, Huang SH, Long LH, Halliwell B (2008) Different patterns of oxidized lipid products in plasma and urine of dengue fever, stroke and Parkinsons disease patients. Cautions in the use of biomarkers of oxidative stress. Antioxid Redox Signal.
114. Soundravally R, Sankar P, Bobby Z, Hoti SL (2008) Oxidative stress in severe dengue viral infection: association of thrombocytopenia with lipid peroxidation. Platelets 19: 447–454.
115. Suvarna JC, Rane PP (2009) Serum lipid profile: a predictor of clinical outcome in dengue infection. Trop Med Int Health 14: 576–585.
116. Marik PE (2006) Dyslipidemia in the critically ill. Crit Care Clin 22: 151–159, viii.
117. Kruger PS (2009) Forget glucose: what about lipids in critical illness? Crit Care Resusc 11: 305–309.
118. Rothwell C, Lebreton A, Young Ng C, Lim JY, Liu W, et al. (2009) Cholesterol biosynthesis modulation regulates dengue viral replication. Virology 389: 8–19.
119. Vaughn DW, Green S, Kalayanarooj S, Innis BL, Nimmannitya S, et al. (2000) Dengue viremia titer, antibody response pattern, and virus serotype correlate with disease severity. J Infect Dis 181: 2–9.
120. Murgue B, Roche C, Chungue E, Deparis X (2000) Prospective study of the duration and magnitude of viraemia in children hospitalised during the 1996–1997 dengue-2 outbreak in French Polynesia. J Med Virol 60: 432–438.
121. Ray G, Kumar V, Kapoor AK, Dutta AK, Batra S (1999) Status of antioxidants and other biochemical abnormalities in children with dengue fever. J Trop Pediatr 45: 4–7.
122. Kalayanarooj S, Nimmannitya S (2005) Is dengue severity related to nutritional status? Southeast Asian J Trop Med Public Health 36: 378–384.
123. Nguyen TH, Nguyen TL, Lei HY, Lin YS, Le BL, et al. (2005) Association between sex, nutritional status, severity of dengue hemorrhagic fever, and immune status in infants with dengue hemorrhagic fever. Am J Trop Med Hyg 72: 370–374.
124. Schaible UE, Kaufmann SH (2007) Malnutrition and infection: complex mechanisms and global impacts. PLoS Med 4: e115.
125. Pacheco P, Bozza FA, Gomes RN, Bozza M, Weller PF, et al. (2002) Lipopolysaccharide-induced leukocyte lipid body formation in vivo: innate immunity elicited intracellular Loci involved in eicosanoid metabolism. J Immunol 169: 6498–6506.
126. de Assis EF, Silva AR, Caiado LF, Marathe GK, Zimmerman GA, et al. (2003) Synergism between platelet-activating factor-like phospholipids and peroxisome proliferator-activated receptor gamma agonists generated during low density lipoprotein oxidation that induces lipid body formation in leukocytes. J Immunol 171: 2090–2098.
127. Castellheim A, Brekke OL, Espevik T, Harboe M, Mollnes TE (2009) Innate immune responses to danger signals in systemic inflammatory response syndrome and sepsis. Scand J Immunol 69: 479–491.
128. Oppenheim JJ, Yang D (2005) Alarmins: chemotactic activators of immune responses. Curr Opin Immunol 17: 359–365.
129. Bianchi ME (2007) DAMPs, PAMPs and alarmins: all we need to know about danger. J Leukoc Biol 81: 1–5.
130. Cinel I, Opal SM (2009) Molecular biology of inflammation and sepsis: a primer. Crit Care Med 37: 291–304.
131. Claus RA, Otto GP, Deigner HP, Bauer M (2010) Approaching clinical reality: markers for monitoring systemic inflammation and sepsis. Curr Mol Med 10: 227–235.
132. Gill R, Tsung A, Billiar T (2010) Linking oxidative stress to inflammation: Toll-like receptors. Free Radic Biol Med 48: 1121–1132.
133. Mockenhaupt FP, Cramer JP, Hamann L, Stegemann MS, Eckert J, et al. (2006) Toll-like receptor (TLR) polymorphisms in African children: Common TLR-4 variants predispose to severe malaria. Proc Natl Acad Sci U S A 103: 177–182.
134. O'Neill LA (2003) Therapeutic targeting of Toll-like receptors for inflammatory and infectious diseases. Curr Opin Pharmacol 3: 396–403.
135. Kim KD, Zhao J, Auh S, Yang X, Du P, et al. (2007) Adaptive immune cells temper initial innate responses. Nat Med 13: 1248–1252.
136. Barton GM (2008) A calculated response: control of inflammation by the innate immune system. J Clin Invest 118: 413–420.
137. Zhao J, Kim KD, Yang X, Auh S, Fu YX, et al. (2008) Hyper innate responses in neonates lead to increased morbidity and mortality after infection. Proc Natl Acad Sci U S A 105: 7528–7533.

Une conception minutieuse de l'étude et de son plan expérimental, conjuguée à la disponibilité d'outils d'analyse, nous ont permis de mettre au point cette approche transcriptomique par la technologie des puces à ADN et d'obtenir de solides résultats. En effet, nous avons non seulement identifié une signature moléculaire large et robuste associée au phénotype DSS, mais aussi mis en évidence la complexité de la réponse de l'hôte au cours de ce syndrome mortel.

Pour cela, une analyse multifactorielle de variance (multi-ways ANOVA) a été utilisée via le logiciel GeneANOVA. Ce type d'analyse permet de mettre en évidence d'éventuelles corrélations parmi les différents paramètres cliniques disponibles mais également de prendre en compte les biais dus à l'expérience (différents temps d'hybridation, de marquage, d'extraction). Ainsi, 2 959 gènes ont pu être identifiés comme étant différentiellement exprimés entre les patients DSS ou DF et DHF, avec un *False Discovery Rate* (FDR) de 10%. La pertinence de la majorité de ces gènes a ensuite été confirmée par une autre approche (SAM) utilisant le logiciel TmeV.

Par la suite, les outils d'analyse ontologique David knowledgebase et *Ingenuity Pathways Analysis* (IPA) nous ont permis de conclure à un enrichissement de notre liste de gènes candidats sous-exprimés chez les DSS en marqueurs de lymphocytes T et de cellules *Natural Killer* ; ce groupe enrichi étant plus précisément associé aux voies de signalisation : « *TCR signaling pathways* » et « *IFN-I related pathways* ». En outre, chez les DSS, nous avons pu observer une augmentation du niveau d'expression de plusieurs marqueurs impliqués dans (1) la réponse anti-inflammatoire, (2) la réparation des tissus, (3) la réponse du complexe pro-inflammatoire et (4) le métabolisme lipidique. L'équipe du Dr. Patricia Paris a ainsi pu suggérer que les mécanismes identifiés seraient fortement impliqués dans la fuite vasculaire massive liée au syndrome DSS.

Les données sont accessibles dans la base de données GEO avec l'identifiant : GSE17924.

3.4.2 Dysautonomie Familiale

La dérégulation de l'épissage des ARNm est un processus crucial dans le développement des maladies neurologiques. La dysautonomie familiale (DF), une maladie orpheline et neurodégénérative, en est un parfait exemple. La DF, aussi appelée syndrome de Riley-Day, est un désordre du système nerveux affectant la survie des neurones dans le système nerveux autonome et sensoriel. Elle affecte presque exclusivement la population juive d'Europe de l'Est avec une incidence annuelle de 1 sur 3 600 naissances. Elle touche aussi bien les hommes que les femmes, dès la naissance, et elle est progressive. Sa transmission est autosomique récessive.

La DF est due à des mutations au sein du gène IKBKAP, localisé sur le bras long du chromosome 9 (9q31). Dans cette maladie, le simple changement T>C en position 6 du site 5' d'épissage (5'ss) de l'exon 20 du gène IKBKAP est responsable de sa non-inclusion au cours de l'épissage du pré-ARNm (Figure 3.9). Cette non-inclusion n'est cependant pas systématique et l'on observe ainsi un épissage alternatif de l'exon 20 avec, chez les patients atteints de DF,

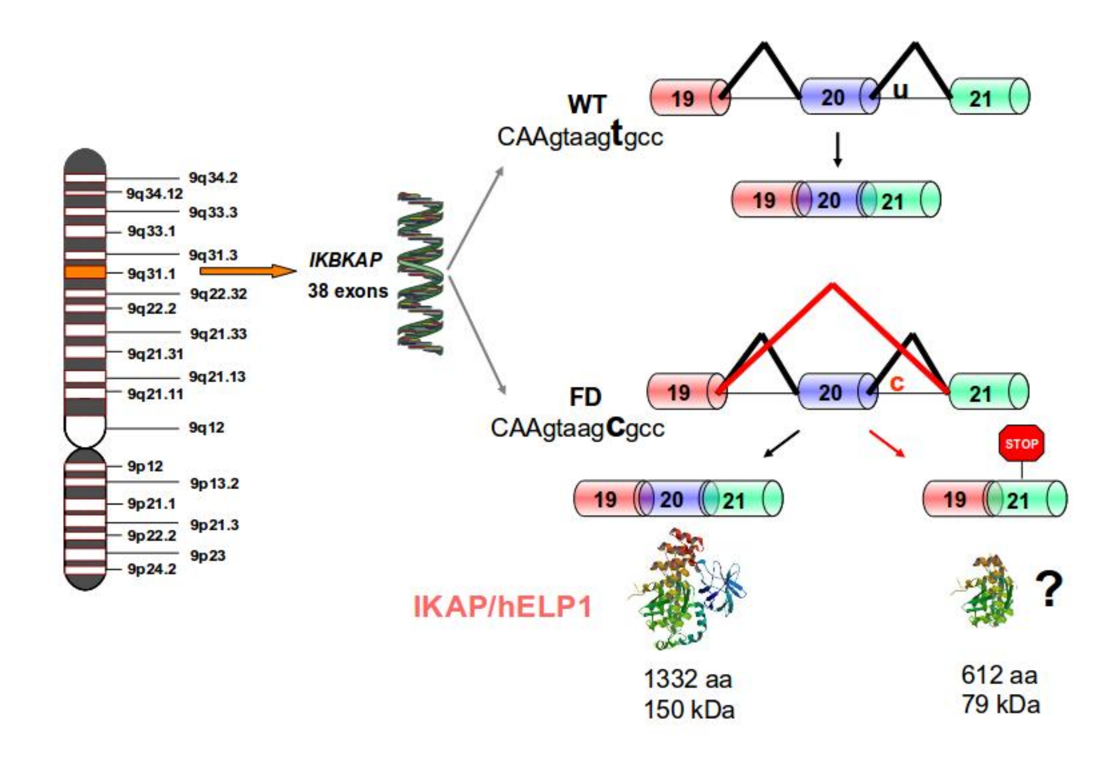


FIGURE 3.9 – Conséquence de l'épissage alternatif du gène IKBKAP sur les différentes isoformes protéique codées par ce gène

une prépondérance de transcrits d'IKBKAP excluant l'exon 20 uniquement dans le système nerveux. La protéine IKAP/hELP1, synthétisée à partir des transcrits incluant l'exon 20, doit jouer un rôle important au niveau du système nerveux, rôle qui reste néanmoins encore très obscur. Le transcrit d'IKBKAP excluant l'exon 20 coderait pour une protéine tronquée de près de 50% du côté C-terminal. Cependant l'existence de cette protéine reste incertaine.

Pour comprendre les voies moléculaires dont l'altération dans le système nerveux cause la DF, nous avons exploré la signature transcriptionnelle de cette maladie. A cette fin, le groupe du Dr. El Chérif Ibrahim a établi des cultures de cellules souches olfactives indifférenciées (hOE-MSC signifiant « *human Olfactory Ecto-Mesenchymal Stem Cells* ») comme modèle d'étude de la DF. Ces cellules contribuent en permanence aux processus de prolifération, migration, différenciation, apoptose et survie cellulaire qui caractérisent la neurogenèse.
Une banque de cellules souches nasales humaines a été établie à partir de 10 individus contrôles et 6 patients DF, permettant de produire des cellules neurales différenciées, à savoir des neurones et des astrocytes. A partir de 5 individus contrôles et 4 patients DF, des cultures de ces cellules souches en cours de différenciation ont été suivies à différents temps : 1, 2, 5 et 9 semaines. Pour chacun de ces temps et de ces échantillons, les ARN totaux extraits de ces cultures ont été amplifiés et marqués radioactivement avant d'être hybridés sur des puces à ADN sur support nylon afin de mettre en évidence les gènes différentiellement exprimés chez les patients DF. Ces puces à ADN comportent 8780 sondes et ont permis d'identifier 46 gènes surexprimés et 4 gènes sousexprimés chez les DF, parmi lesquels 10 gènes avaient précédemment été cités par des études indépendantes.

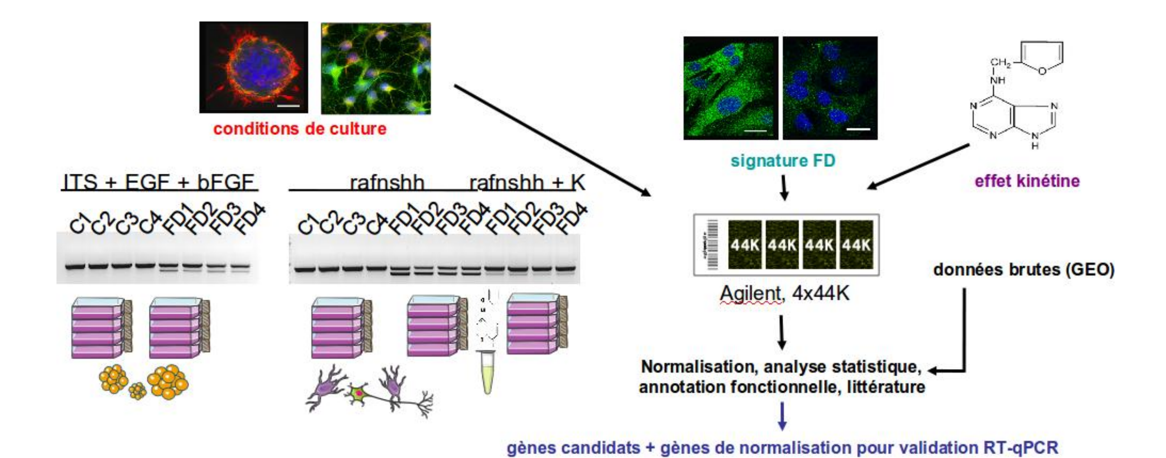


FIGURE 3.10 – Résumé du plan expérimental et analytique de la seconde campagne de puces à ADN

Cette première analyse a fait l'objet d'une publication dans PLoS ONE en 2010. Les données ont été publiées sous ArrayExpress avec l'identifiant : E-MTAB-281.

Afin de confirmer les résultats précédents et d'identifier de nouvelles cibles thérapeutiques potentielles à l'échelle pan-génomique cette fois-ci, une nouvelle campagne de puces à ADN a été réalisée sur puces pan-génomiques Agilent™à partir de 4 DF et de 4 contrôles à l'état de sphères mais également à l'état différencié (Figure 3.10). Au cours de cette étude, j'ai réalisé la partie de traitement et d'analyse des données et ainsi que leur soumission à GEO. Dans la mesure où il n'existe aucun traitement spécifique à la DF permettant de réduire les symptômes ni même de contrecarrer l'avancement de la maladie, des recherches ont été menées pour tester une molécule prometteuse : la kinétine. Le choix s'est porté sur cette molécule car elle permet de corriger l'épissage alternatif aberrant d'IKBKAP bien que son mécanisme d'action soit totalement inconnu ([Boone *et al.*, 2010], [Hims *et al.*, 2007], [Keren *et al.*, 2010], [Lee & Mahadevan, 2009], [Slaugenhaupt *et al.*, 2004]). Ainsi, nous avons recherché sa signature transcriptionnelle afin de mieux comprendre son mode d'action. La signature transcriptionnelle des DF obtenue a ensuite été comparée aux données brutes de 5 autres études publiées et accessibles sur GEO ou ArrayExpress.

A l'aide d'un SAM (FDR 0%), 3000 gènes ont été trouvés comme étant différentiellement exprimés entre les cellules souches indifférenciées (sphères) et les cellules neuro-gliales différenciées (neurones et astrocytes). Nous avons ainsi pu clairement valider l'empreinte transcriptionnelle induite par les 3 facteurs du cocktail de différenciation neuro-gliale utilisé lors de la précédente campagne de puces à ADN en retrouvant des gènes connus pour leur réponse à (1) l'acide rétinoïque, (2) la forskoline et (3) le morphogène Sonic hedgehog. 37 gènes ont été identifiés par SAM (FDR 10%) comme permettant de distinguer toutes les cultures de DF par rapport à celles des contrôles avec une prépondérance de gènes ayant un rôle déterminant dans le fonctionnement du système nerveux. La comparaison de notre étude avec les 5 transcriptomes provenant de GEO, nous a permis d'identifier une centaine de gènes dont les variations d'expression entre échantillons contrôles et DF (ou knock down pour IKAP/hELP1) sont conservées dans au moins deux études indépendantes. Parmi les processus

qui semblent altérés de façon récurrente dans la DF, nous avons pu identifier : la différenciation neuronale, la migration et l'adhésion cellulaires, et la régulation de l'apoptose. Enfin, parmi les gènes dérégulés par la kinétine, nous avons pu mettre en évidence pour la première fois deux facteurs d'épissage participant à la reconnaissance du site 5' d'épissage (5'ss). Cela ouvre ainsi de nouvelles pistes pour le décryptage du mode d'action de la kinétine sur l'épissage du pré-ARNm d'IKBKAP.

Tous ces travaux font l'objet d'une publication sous presse à Human Mutation. Les données sources sont accessibles sur GEO à l'aide de l'identifiant : GSE27915.



Olfactory Stem Cells, a New Cellular Model for Studying Molecular Mechanisms Underlying Familial Dysautonomia

Nathalie Boone[1], Béatrice Loriod[2], Aurélie Bergon[2], Oualid Sbai[1], Christine Formisano-Tréziny[3], Jean Gabert[3,4], Michel Khrestchatisky[1], Catherine Nguyen[2], François Féron[1], Felicia B. Axelrod[5], El Chérif Ibrahim[1]*

1 NICN-CNRS UMR 6184, Université de la Méditerranée-Faculté de Médecine Nord, IFR Jean Roche, Marseille, France, **2** TAGC, INSERM U928, Marseille, France, **3** Plateforme Transcriptome, CRO2, Faculté de Médecine, Marseille, France, **4** Biochemistry and Molecular Biology, Hôpital Nord, AP-HM, Marseille, France, **5** Department of Pediatrics, New York University School of Medicine, New York, New York, United States of America

Abstract

Background: Familial dysautonomia (FD) is a hereditary neuropathy caused by mutations in the *IKBKAP* gene, the most common of which results in variable tissue-specific mRNA splicing with skipping of exon 20. Defective splicing is especially severe in nervous tissue, leading to incomplete development and progressive degeneration of sensory and autonomic neurons. The specificity of neuron loss in FD is poorly understood due to the lack of an appropriate model system. To better understand and modelize the molecular mechanisms of *IKBKAP* mRNA splicing, we collected human olfactory ecto-mesenchymal stem cells (hOE-MSC) from FD patients. hOE-MSCs have a pluripotent ability to differentiate into various cell lineages, including neurons and glial cells.

Methodology/Principal Findings: We confirmed *IKBKAP* mRNA alternative splicing in FD hOE-MSCs and identified 2 novel spliced isoforms also present in control cells. We observed a significant lower expression of both *IKBKAP* transcript and IKAP/hELP1 protein in FD cells resulting from the degradation of the transcript isoform skipping exon 20. We localized IKAP/hELP1 in different cell compartments, including the nucleus, which supports multiple roles for that protein. We also investigated cellular pathways altered in FD, at the genome-wide level, and confirmed that cell migration and cytoskeleton reorganization were among the processes altered in FD. Indeed, FD hOE-MSCs exhibit impaired migration compared to control cells. Moreover, we showed that kinetin improved exon 20 inclusion and restores a normal level of IKAP/hELP1 in FD hOE-MSCs. Furthermore, we were able to modify the *IKBKAP* splicing ratio in FD hOE-MSCs, increasing or reducing the WT (exon 20 inclusion):MU (exon 20 skipping) ratio respectively, either by producing free-floating spheres, or by inducing cells into neural differentiation.

Conclusions/Significance: hOE-MSCs isolated from FD patients represent a new approach for modeling FD to better understand genetic expression and possible therapeutic approaches. This model could also be applied to other neurological genetic diseases.

Citation: Boone N, Loriod B, Bergon A, Sbai O, Formisano-Tréziny C, et al. (2010) Olfactory Stem Cells, a New Cellular Model for Studying Molecular Mechanisms Underlying Familial Dysautonomia. PLoS ONE 5(12): e15590. doi:10.1371/journal.pone.0015590

Editor: Carlo Gaetano, Istituto Dermopatico dell'Immacolata, Italy

Received September 2, 2010; **Accepted** November 13, 2010; **Published** December 20, 2010

Funding: The authors thank the Association Française de Recherche contre les Myopathies (AFM) for supporting their work. NB was supported by a PhD fellowship from the Ministère de l'Education Nationale, de la Recherche et de la Technologie (MENRT). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: el-cherif.ibrahim@univmed.fr

Introduction

Familial dysautonomia (FD, Riley-Day syndrome, hereditary sensory and autonomic neuropathy type III, MIM 223900) is an autosomal recessive genetic disorder that occurs in 1:3600 live births with a carrier frequency of 1 in 30 in the Ashkenazi Jewish population. The disease is characterized by incomplete development and the progressive depletion of autonomic and sensory neurons [1–3] resulting in variable symptoms including: insensitivity to pain, lack of overflow tearing, inappropriate blood pressure control manifested as orthostatic hypotension and episodic hypertension, poor oral coordination resulting in poor feeding and swallowing, and gastrointestinal dysmotility [4]. No cure is available for this disorder and treatment is aimed at controlling symptoms and avoiding complications.

FD is caused by mutations in the *IKBKAP* gene which encodes a protein termed IKAP/hELP1 [5,6]. The most prevalent mutation, is a splice mutation; the T-to-C transition in position 6 of the 5′ splice site (5′ss) of intron 20 (IVS20$^{+6T\rightarrow C}$) of this gene. All FD cases have at least one copy of this mutation; >99.5% are homozygous [5–7]. This mutation leads to variable, tissue-specific skipping of exon 20 of *IKBKAP* mRNA, with the central and peripheral nervous system more prone to complete skipping than others tissues, which leads to reduced IKAP/hELP1 protein levels [8].

Although the exact function of the IKAP/hELP1 protein is not clearly understood, researchers have identified IKAP/hELP1 as the scaffold protein required to assemble a well conserved six-protein complex (ELP1-6) called the holo-Elongator complex that possess histone acetyltransferase activity directed against histone H3 and H4 *in vitro* [9]. IKAP/hElongator is recruited to the transcribed regions of some human genes essentially involved in actin cytoskeleton regulation and cell motility migration [10]. This role may underlie a cell motility deficiency in FD neurons because of impaired transcriptional elongation of some genes coding for proteins involved in cell migration. Indeed, one study found that mouse neurons defective in Elongator exhibit reduced levels of acetylated α-tubulin, causing defects in radial migration and branching of cortical projections neurons [11]. Another study showed that *Caenorhabditis elegans* Elongator complex is required for correct acetylation of microtubules and neuronal development [12]. IKAP/hELP1 protein is also involved in other cellular processes, including tRNA modifications [13–15], exocytosis [16], and zygotic paternal genome demethylation [17]. Recently, its homolog in fly (D-elp1) has also been suggested to be involved in RNA interference through a RNA-dependent RNA polymerase activity [18].

To better understand the molecular mechanisms leading to aberrant splicing of *IKBKAP* mRNA in FD, creation of model systems recapitulating the pathological development of neural cells is required. Because *IKBKAP* gene knock out causes embryonic lethality [19], an animal model that exhibits the major phenotypic characteristics observed in FD humans has not yet been established. However, a humanized *IKBKAP* transgenic mouse model for FD has been created [20], that reproduces the tissue-specific splicing of *IKBKAP* mRNA in nervous tissues. Such a model is a notable progress in the comprehension of this complex rare disease and offers a potential system for testing therapeutic agents. However, transgenic animals do not reproduce phenotypic features of FD as they maintain normal development. Alternatively, FD patient fibroblasts are an informative model of mRNA splicing regulation. However, a recent study suggests that IKAP/hELP1 expression is much higher in neurons compared to fibroblasts [21], and fibroblasts do not exhibit the same ratio of *IKBKAP* exon 20 including:exon 20 skipping transcripts (named WT:MU respectively for simplicity) as observed in nervous system-derived tissues [8]. This finding narrows the understanding of disease mechanisms in a neural context. Finally, generation of neural cells through the production of induced pluripotent stem (iPS) cell from FD fibroblasts has been recently established [22]. Neural cells derived from iPS cells have potential to be used for studies of neuropathologies [23]. However, the labor intensive reprogramming required to induce iPS cells erases the developmentally relevant epigenetic signature specific to the disease state. As a consequence, some important information may be lost impeding recreation of an accurate disease model. The demonstration that fibroblasts can be converted directly into neurons, without an initial reprogramming, as recently evidenced in mouse [24], is very attractive. Nevertheless, during their reprogramming, human iPS cells do not pass through the normal stages of embryonic development that human ES cells undergo. Although both stem cell types share a common transcriptional signature, a subset of genetic profiles found in human iPS cells suggests retention of transcriptional and epigenetic memory related to their tissue of origin, which can substantially affect their potential to differentiate into different cell types [25–27]. Thus, cells collected from primary sources that have been subjected to environmental signals appropriate for the pathological specificity of the targeted disease are likely important to mirror the biology of diseased human neural cells.

Our aim is to understand what mechanisms drive *IKBKAP* mRNA splicing to the almost exclusive production of aberrant transcripts (MU) in neuronal cells. Here we demonstrate the potential of human olfactory ecto-mesenchymal stem cells (hOE-MSCs) to model this aspect in FD. Indeed, neurogenesis occurs throughout adult life in the olfactory mucosa, due to the presence of resident multipotent stem cells giving rise to olfactory neurons *in vivo* [28]. hOE-MSCs can be grown into neurospheres, that are multipotent, and differentiate *in vitro* into neurons, astrocytes, and oligodendrocytes as well as other cell types [29,30]. Isolated from patients, cultures of hOE-MSCs provide potential models for genetically determined neuropsychiatric diseases [31–33], and stand as an interesting human model to investigate gene networks and cellular pathways altered in disease like FD. For example, cell migration defects have been observed in cells lacking normal expression of IKAP/hELP1 [10,11,22,34], and we show here that FD hOE-MSCs exhibit impaired migration compared to control cells. Additionally, hOE-MSCs are an appropriate model for validating the potency of therapeutic agents such as kinetin, a cytokinin that has been shown to increase *IKBKAP* mRNA and protein expression in FD cell lines and *in vivo* models [20,22,35,36] as well as in leukocytes of healthy carriers of the FD mutation [37].

Results

FD hOE-MSCs express stem cell, glial and immature neuronal markers

To establish a human cellular model of FD, we collected 4 olfactory mucosa biopsies from homozygous patients for the $IVS20^{+6T\rightarrow C}$ FD mutation. As previously demonstrated with control biopsies [30], after about 2 weeks of culture, the microscopic examination of the tissue crushed under a glass coverslip revealed stem cell proliferation (Figure 1A and 1B). After reaching confluency in a 4-well plate, the cells attached to the glass coverslip were further expanded by transfer into a 6-well plate (Figure 1C). Like control hOE-MSCs, we observed that FD hOE-MSCs could be cultured for long periods (at least 15 cycles of trypsin/EDTA treatment and expansion on larger plastic surface) with a doubling time of about 30–48 h. When subjected to immunostaining, all hOE-MSCs derived from control and FD biopsies express the neural stem cell–specific marker nestin (Figure 1D and 1E) and the immature neuronal marker β-III tubulin (Figure 1F and 1G) in the same proportions (Figure 1H and 1I). A comparatively low GFAP staining was observed in every hOE-MSCs (Figure 1J and 1K). In addition, cells were negative for a mature neuronal marker, MAP2 (Figure 1L and 1M). This analysis suggests that both control and FD hOE-MSCs display properties of neuroglial progenitor cells.

Expression of *IKBKAP* transcripts is dramatically reduced in FD hOE-MSCs

IKBKAP mRNA expression was investigated in cultures of 5 controls and 4 FD hOE-MSCs at early (P1, P2) and later cell passages (P5, P9). A semi quantitative RT-PCR analysis revealed that, while control hOE-MSCs expressed exclusively the WT mRNA transcript (Figure 2A, left panel), FD hOE-MSCs expressed the WT but also the MU transcript (Figure 2A, right panel). We also demonstrated that long time culture conditions and trypsin-EDTA mediated cell passages did not affect the *IKBKAP* gene expression pattern. In order to more accurately determine the level of expression of *IKBKAP* alternative transcripts, we designed primers, probes and plasmid calibrators to perform absolute quantification using quantitative real-time RT-PCR (RT-qPCR) on the same samples. Strikingly, WT transcripts were

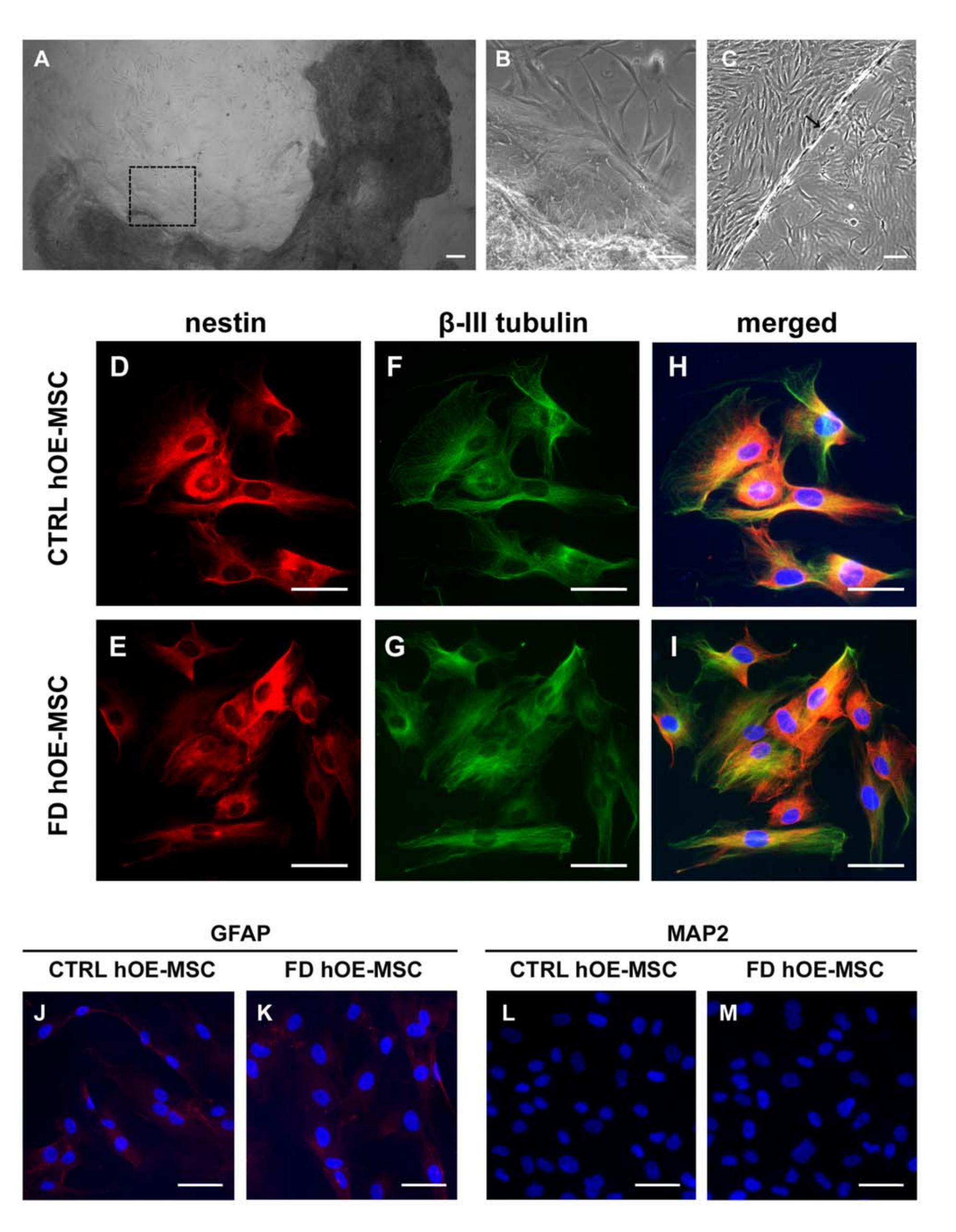


Figure 1. hOE-MSCs display characteristics of immature neuroglial cells. A, lamina propria (dark) from an FD olfactory mucosa biopsy was placed under a glass coverslip to initiate stem cell proliferation. Area delimited by a black square is enlarged in B. C, after transfer in a 6-well plate, cells attached to the coverslip (arrow) proliferated and colonized the complete area of the well. (D–M), Immunofluorescence stainings of both control (CTRL) and FD hOE-MSCs are positive for nestin (D, E), β-III tubulin (F,G) with similar expression levels (H, I) while slightly positive for GFAP (J,K) and negative for MAP2 (L, M). Green represents Alexa fluor-488, red Alexa Fluor-594. Nuclei (blue) were stained with Hoechst blue. Scale bars represent 50 μm.

doi:10.1371/journal.pone.0015590.g001

much less expressed in FD (5–8 fold), when compared to controls hOE-MSCs (Figure 2B). In addition, WT and MU transcripts were present in nearly equal amounts in FD hOE-MSCs (Figure 2B, right graph). Furthermore, the total amount of *IKBKAP* transcripts in FD (WT+MU) remains 3 to 5 times less abundant than WT in controls, which suggests a defect in *IKBKAP* transcription and/or mRNA stability. In FD cells, the differential expression of *IKBKAP* transcripts was also correlated to a reduced expression of IKAP/hELP1 protein in FD, when compared to controls, as revealed by western blot analysis (Figure 2C). Since MU transcripts contain a premature stop codon that may activate the nonsense-mediated mRNA decay (NMD) pathway, we wanted to confirm whether this pathway is responsible for the lower *IKBKAP* transcripts expression in FD cells. Thus, we tested cycloheximide, a protein synthesis inhibitor which also inhibits NMD. Indeed, FD cells preincubated for 6 h with cycloheximide exhibited a stabilization of the MU transcript as evidenced by semi-quantitative RT-PCR (Figure 2D, left panel). To accurately determine the level of WT and MU *IKBKAP* transcripts in these samples, absolute RT-qPCR analysis was performed (Figure 2D, right panel). The results clearly demonstrated that the WT:MU ratio decreases when mRNA surveillance is inhibited. Thus, a

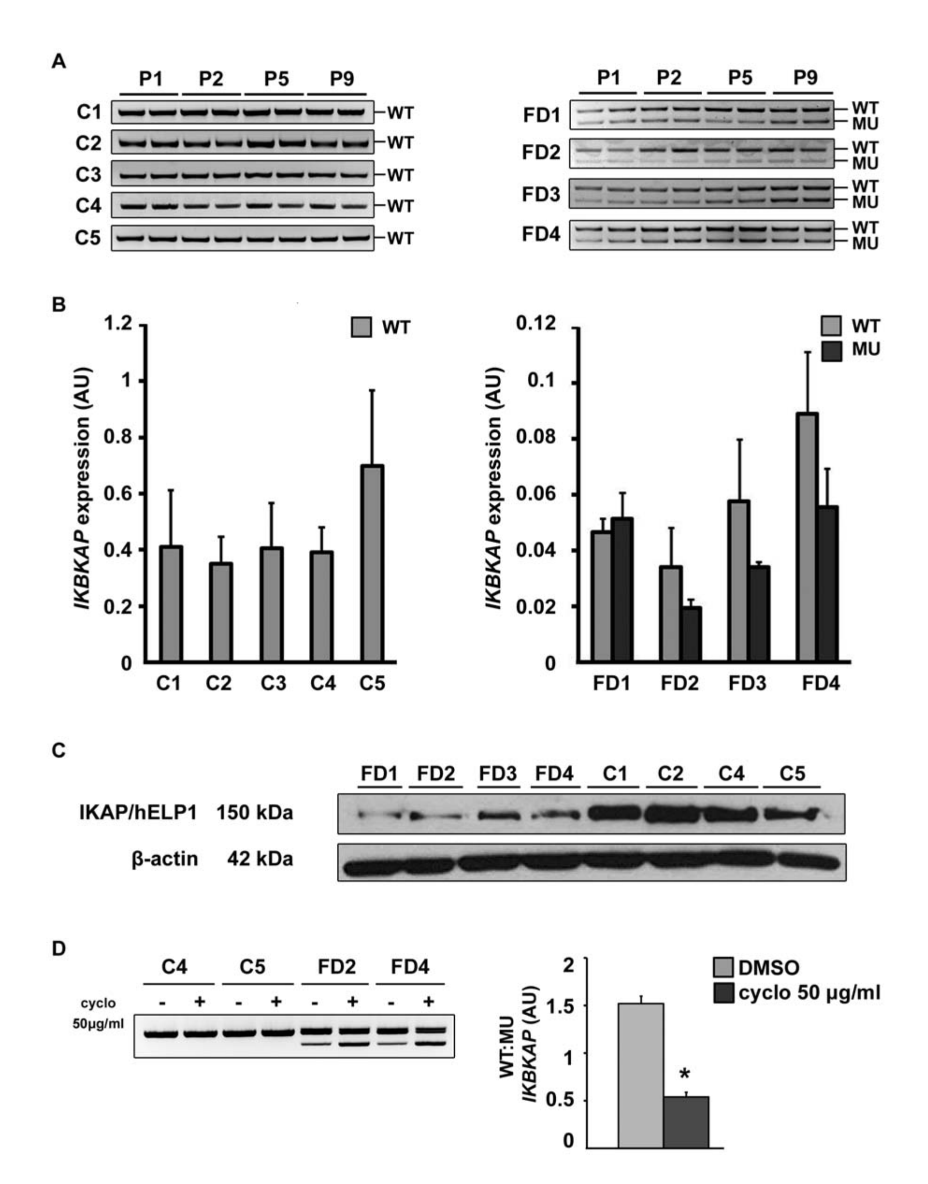


Figure 2. Expression of *IKBKAP* transcripts and IKAP/hELP1 protein in hOE-MSCs. A, agarose gel electrophoresis of end-point RT-PCR products showing both WT and MU transcripts of *IKBKAP* gene for control (left panel) and FD hOE-MSCs (right panel) at cell passage 1,2,5,9. B, graph of the mean level of expression of *IKBKAP* alternative transcripts in control (left panel) and FD hOE-MSCs (right panel) at cell passages 1,2,5,9, determined by absolute RT-qPCR. *ABL1* was used as a reference gene for normalization. Error barrs denote standard error. C, western blot analysis of total lysate from 4 controls and 4 FD hOE-MSCs using monoclonal anti-IKAP/hELP1 antibody (upper panel). Anti-β-actin was included to show equal loading (lower panel). D, NMD pathway was blocked by the translation inhibitor cycloheximide and results in an elevated expression of MU transcripts in FD cells (agarose gel electrophoresis, left panel). Results are confirmed with absolute qPCR normalized with *ABL1* (right panel).
doi:10.1371/journal.pone.0015590.g002

large amount of *IKBKAP* MU transcripts is degraded through the NMD pathway resulting in much less *IKBKAP* transcripts and IKAP/hELP1 protein in FD compared to control cells.

Heterogeneous IKAP/hELP1 distribution in hOE-MSCs

Since the localization of IKAP/hELP1 remains controversial and is important to understand protein functions, we stained both control and FD hOE-MSCs with the monoclonal antibody directed against IKAP/hELP1 and previously used for detecting the protein by western blot analysis. In control cells, confocal imaging revealed a weak and diffuse signal with a dominant cytoplasmic staining within the perinuclear area. We could also detect the presence of IKAP/hELP1 in the nucleus of hOE-MSCs (Figure 3A–C). Significantly, FD hOE-MSCs exhibit a weaker anti-IKAP/hELP1 immunofluorescence staining compared to control cells, with a similar distribution of the staining (Figure 3D–F). Therefore, collectively, our results are in agreement with a wide distribution of IKAP/hELP1, including a much lower IKAP/hELP1 staining in FD hOE-MSCs, in agreement with RT-qPCR and western blot analysis.

Transcriptome analysis identified fifty dysregulated genes

It is widely accepted that culture conditions alone may exert effects on gene expression, resulting in experimental inconsistencies [38,39]. Thus, to investigate the involvement of candidate disease mechanisms in FD and to test whether differences in gene expression are stably imprinted in FD compared to control hOE-

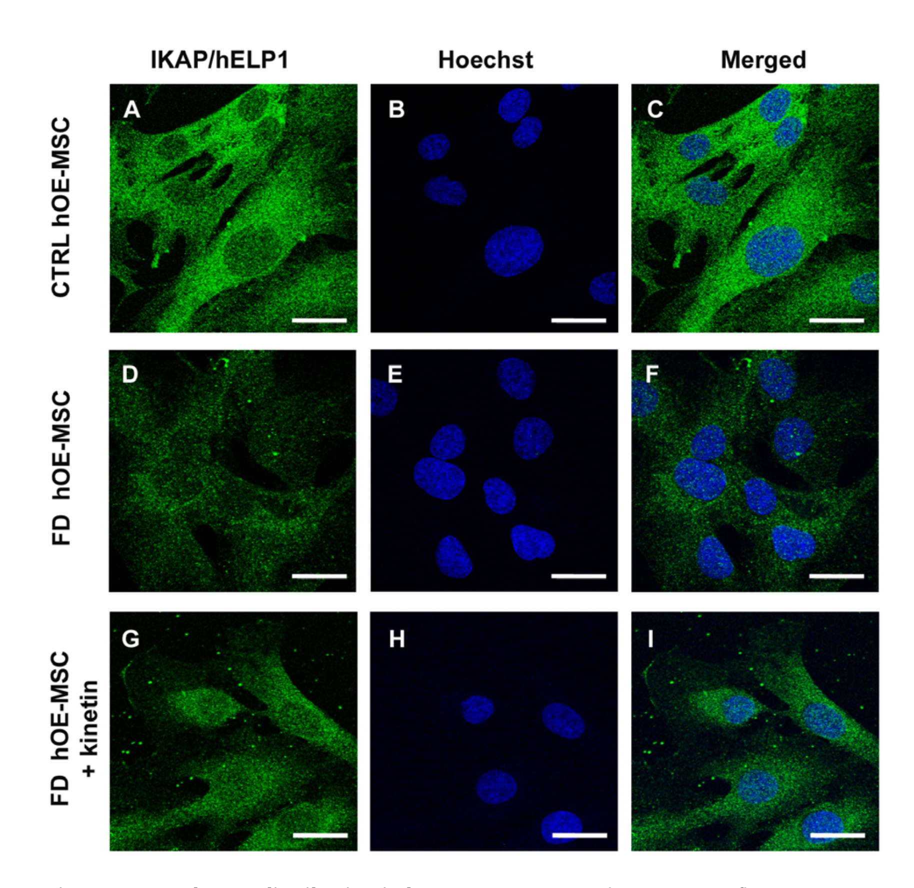


Figure 3. IKAP/hELP1 distribution in hOE-MSCs. Anti-IKAP/hELP1 immunofluorescence staining in control (A, B, C), FD hOE-MSCs (D, E, F), and FD hOE-MSCs treated with 100 μM kinetin for 24h (G, H, I). The primary antibody used is a mouse monoclonal anti-IKAP/hELP1. Scale bars represent 20 μm.
doi:10.1371/journal.pone.0015590.g003

MSCs, we explored the transcriptome of these cells at very early (P1, P2) and later (P5, P9) cell passages with the same samples used to quantify *IKBKAP* transcripts. Among the 8,780 cDNA represented on the microarray, 46 were significantly decreased and only 4 increased in FD hOE-MSCs, when compared to control hOE-MSCs (fold-change>1.4-fold; p-value<6.10^3, Table 1 and Table S1), considering a false discovery rate (FDR) of 3% (Figure S1). Notably, the biological processes and the signaling pathways most significantly targeted by the effectors on our list were actin cytoskeleton organization, cell growth, and apoptosis (Table 1). More specifically, we identified 10 genes (Table 1 and Table S1) that also exhibited a significant dysregulated expression in previous microarray studies [10,22]. Interestingly, 2 genes, *PMEPA1* and *GSN* (encoding TMEPAI and gelsolin, respectively), involved in cell growth and cytoskeleton organization, respectively, were dysregulated in both the *IKBKAP* RNAi and FD iPS cell studies.

In order to assess the robustness of our microarray analysis, RT-qPCR analysis was performed, on independent RNAs extracted from 4 control and 4 FD hOE-MSCs harvested at the second, fourth, and seventh cell passage. Since gene expression quantification using RT-qPCR requires a steady reference gene, we selected three genes frequently used for normalization of the data, *ABL1*, *RPLP0*, and *HPRT1*. We confirmed that *PMEPA1* (Figure 4A), the most dysregulated gene on the microarray, and *S100A16* (Figure 4B), were significantly underexpressed in FD samples. The expression pattern of these two candidate genes was essentially identical at all passages with the 3 reference genes, which demonstrates the validity and reliability of the array data.

FD OE-MSCs migration is altered compared to controls

To explore the functional consequence of a down-regulated expression of genes involved in cell migration in FD hOE-MSCs compared to control cells, we used the Boyden's chamber assay. After comparing the migration pattern of 3 control and 3 FD hOE-MSCs in serum medium and serum-free medium (ITS), we determined that FD cells invasion is significantly reduced compared to control cells both in serum and in ITS medium (Figure 5).

Confirmed down-expression of first and final *IKBKAP* exons in FD hOE-MSCs

Since we and others [22,40] did not detect *IKBKAP* among the significantly down-regulated transcripts in FD compared to control samples, we asked whether this discrepancy could be due to a lack of sensitivity of microarray compared to RT-qPCR. For this purpose, we decided to analyze *IKBKAP* levels of expression by investigating other exons distal from *IKBKAP* exon 20. By looking at the beginning of *IKBKAP* transcript, we identified a second event of alternative splicing. After amplifying transcripts from exon 1 to exon 5, we obtained 2 PCR products (Figure 6B, upper panel). The sequencing of the less abundant and shorter PCR product revealed the use of an alternative 3′ss within *IKBKAP* exon 2, which is shortened of 145 nt (Figure 6A, left schematic). Accordingly, the loss of the ATG start codon located within the 5′ end of exon 2 can potentially induce the use of an alternative ATG start codon (in exon 4), resulting in the synthesis of a putative 114 amino acid-truncated IKAP/hELP1 protein (Figure 6D).

Table 1. Most dysregulated genes in FD hOE-MSCs are involved in seven over-represented cellular processes.

Gene	Clone ID	FC	*p*-value	Biological process	Studies
				Actin cytoskeleton reorganization *p* = 0.000275	
GSN	246170	−1.91	0.00124	actin cytoskeleton reorganization	*1 *2
MYO9B	279085	−1.70	0.00144	actin cytoskeleton remodeling	
MYPN	325601	−1.52	0.00131	sarcomere organization through nebulin and α-actinin interactions	
DSTN	149199	−1.48	0.00017	actin filament depolymerization	
CORO2B	547561	−1.44	0.00079	neuronal actin structure reorganization	
SLC9A3R2	155467	2.42	0.00034	adaptor of ion channels and receptors to the actin cytoskeleton	*1
				Regulation of apoptosis *p* = 0.00203	
CHEK2	1893020	−2.20	0.00007	cell cycle arrest and apoptosis in response to DNA damage	*1
ZMAT3	525407	−1.78	0.00251	positive regulation of p-53-mediated apoptosis	
TNFSF10	713945	−1.75	0.00041	induction of apoptosis by activation of caspase activity	*2
PARP3	436086	−1.60	0.00004	positive regulation of apoptosis- maintenance of genomic stability	
				Transport *p* = 0.00224	
ABCG5	121977	−1.59	0.00089	cholesterol transport in and out of the enterocytes	
SLC35E1	487960	−1.55	0.00086	monosaccharide transport	
SLC22A6	36482	−1.52	0.00050	α-ketoglutarate transmembrane transporter activity	
SFXN2	757192	−1.45	0.00021	iron transport	
				Cell proliferation *p* = 0.00552	
APOE	1870594	−1.68	0.00028	cell proliferation-regulation of neurite extension	*1
CD22	284220	−1.58	0.00032	B cell proliferation	
CD38	123264	−1.54	0.00074	B cell proliferation	
GBA	757264	−1.53	0.00012	cell proliferation-ceramide metabolic process	
SERINC2	149995	1.84	0.00014	cell proliferation	
				Regulation of cell growth and cell cycle *p* = 0,0091	
PMEPA1	366599	−4.92	0.00585	EGF receptor signaling pathway - negative regulation of cell growth	*1 *2
STRBP	669157	−1.75	0.00007	regulation of cell growth	
INO80B	323554	−1.63	0.00104	growth induction and cell cycle arrest at the G1 phase	
S100A16	739851	−1.55	0.00012	regulation of cell cycle progression	*1
CDIPT	306047	−1.49	0.00088	regulation of cell growth	
				Nervous system process *p* = 0,0302	
LRCH1	683580	−3.62	0.00080	long-term memory and learning - signal transduction	
KCNT2	38677	−2.06	0.00010	synaptic transmission mediated by K (+) channels	
NUMBL	1855110	−1.61	0.00118	Notch signaling pathway inhibition - cerebral cortex morphogenesis	
DULLARD	346368	−1.53	0.00198	Nuclear organization-negative regulation of BMP signaling	
				Proteolysis *p* = 0,0334	
FBXL15	166240	−1.71	0.00130	ubiquitin-dependent protein catabolic process	
WSB1	298983	−1.61	0.00123	ubiquitination and proteosomal degradation of target proteins	
PCSK7	241130	−1.53	0.00155	proteolysis- ubiquitous endoprotease activity	
RNF115	471834	−1.42	0.00067	proteolysis- vesicle-mediated transport - vesicle traffic	
MMP27	767086	−1.35	0.00067	proteolysis of fibronectin, laminin, gelatins and/or collagens	

Clone ID represents the number assigned to the original clones produced by the I.M.A.G.E Consortium. FC = Fold change, and *p*-values were calculated by SAM analysis as described in Methods. This list of genes was annotated with the Explain™ System from Biobase. 7 majors processes are overrepresented in our list of genes, and, for each process, *p*-values were calculated and adjusted by the Bonferroni correction. The last column indicates the genes that were also found to be significantly dysregulated in 2 previous FD studies.
*1 = Lee et al. 2009.
*2 = Close et al. 2006.
doi:10.1371/journal.pone.0015590.t001

When investigating expression at the end of *IKBKAP* coding sequence, again we observed a third alternative splicing event. The amplification from exon 33 to exon previously numbered exon 36 (and now called exon 37) revealed 2 products (Figure 6B, middle panel). The sequencing of the barely detectable and longer PCR product revealed the inclusion of an additional exon (Figure 6A, right schematic). This exon inclusion also induced a frameshift and resulted in a premature stop codon whose relative location may

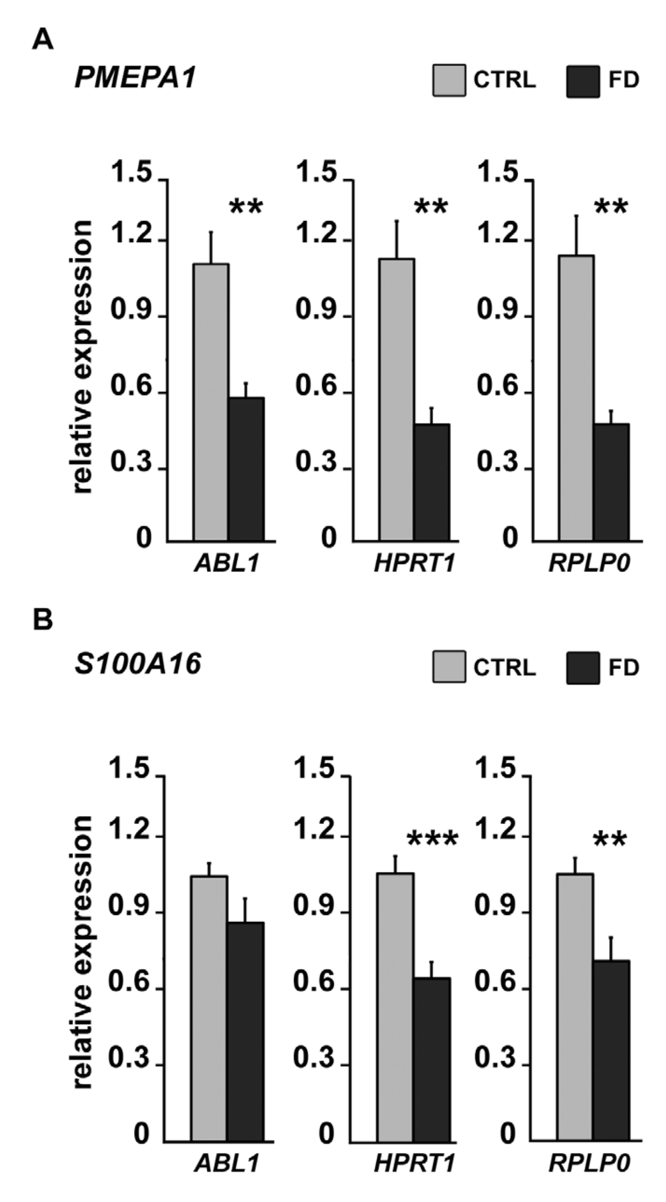


Figure 4. Relative levels of expression of *PMEPA1* and *S100A16* transcripts determined by RT-qPCR. RT-qPCR using total RNAs extracted from 4 controls and 4 FD hOE-MSCs at cell passages 2, 4 and 7. Histograms represent the mean value of *PMEPA1* (A) and *S100A16* (B) transcript expression level, relative to 3 reference genes *ABL1*, *HPRT1*, *and RPLP0* in control (grey) and FD samples (black). Error bars denote standard errors. (* $P<0.05$; ** $P<0{,}01$, *** $P<0{,}001$ using two-tailed Student's test).
doi:10.1371/journal.pone.0015590.g004

lead to NMD of this new isoform (Figure 6D). We confirmed this exon 36 inclusion with specific primers (Figure 6B, lower panel).

Both new alternative splicing events we described were also observed in others cell types (fibroblasts, HeLa, peripheral mononuclear cells, data not shown) and we decided to focus on the two major splicing events, full exon 2 inclusion and exon 36 skipping. We derived the tools (plasmids, primers, probes) to perform absolute quantification of full-length exon 2 inclusion and exon 36 skipping by RT-qPCR on samples from 4 controls and 4 FD hOE-MSC cultures. Similar underexpression of *IKBKAP* transcripts (WT+MU) was observed in FD cells compared to control cells, regardless of the exon investigated (Figure 6C). Thus,

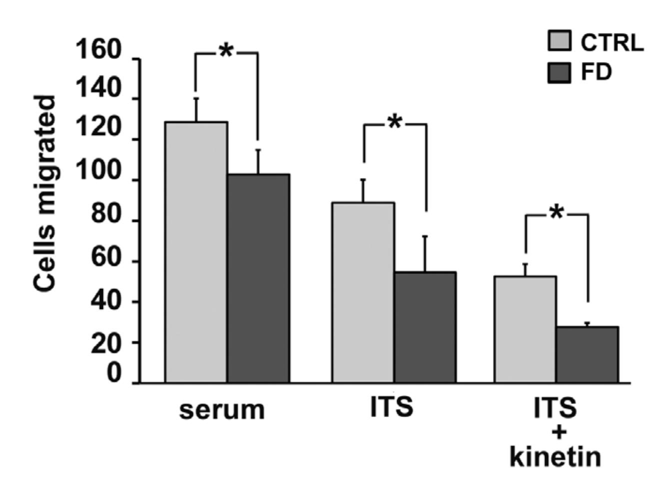


Figure 5. FD hOE-MSCs demonstrate reduced migration. Cell invasion in 3 different controls and 3 different FD hOE-MSCs was studied using Boyden chamber assay. Cells (3×10^4) were added to the upper chamber in serum medium, serum-free medium (ITS), or ITS supplemented with 100 µM kinetin. Cell invasion was mesured after 24 h. Results are shown as the average ± SEM of the number of cells per microscopic field. (* $P<0.05$).
doi:10.1371/journal.pone.0015590.g005

these results confirmed a decreased level of *IKBKAP* transcripts (WT+MU) in FD cells.

In addition, we tested for the stability of exon 36-containing transcripts after cycloheximide treatment. Absolute RT-qPCR analysis revealed that the exon 36 skipping : exon 36 inclusion ratio decreases when NMD pathway is inhibited (Figure S2), suggesting that transcripts including exon 36 are degraded through NMD in FD, as well as in control OE-MSCs (data not shown).

Kinetin treatment corrects aberrant *IKBKAP* pre-mRNA splicing

Our next goal was to assess whether the production of both WT and MU *IKBKAP* mRNAs can be modulated in our model. In previous studies, one compound, kinetin (6-furfurylaminopurine) was found to correct *IKBKAP* splicing and increase IKAP/hELP1 production in FD cells [35]. We tested whether this drug could also modify the splicing defect of *IKBKAP* in FD hOE-MSC cells. For this purpose, we used increasing concentrations of kinetin (25 to 200 µM) on a FD hOE-MSCs culture for 72 h. As expected, after semi-quantitative RT-PCR, we observed a significant decrease of MU transcript compared to non-treated cells on agarose gel electrophoresis (Figure 7A). The level of *IKBKAP* mRNA splicing correction increased proportionally to the concentration of kinetin, and the MU transcript almost vanished at 100 µM. The dose-dependent action of kinetin on increasing WT:MU ratio was confirmed by RT-qPCR analysis (Figure 7B). A similar finding was observed when IKAP/hELP1 proteins were detected by western blot analysis (Figure 7C). Accordingly, when FD hOE-MSC were incubated with 100 µM kinetin for 24 h, we observed a major increase of anti-IKAP/hELP1 staining, in cytoplasmic as well as in nuclear areas (Figure 3G–I). However, the same kinetin treatment could not rescue the migration defect observed in FD hOE-MSCs with the Boyden's chamber assay (Figure 5).

In order to determine how fast kinetin modulates *IKBKAP* mRNA splicing, we performed a time-course experiment with a constant concentration of 80 µM over 24 h. After performing semi-quantitative RT-PCR analysis, the first significant increase of WT:MU ratio was seen after 24 h of kinetin treatment (Figure 7D).

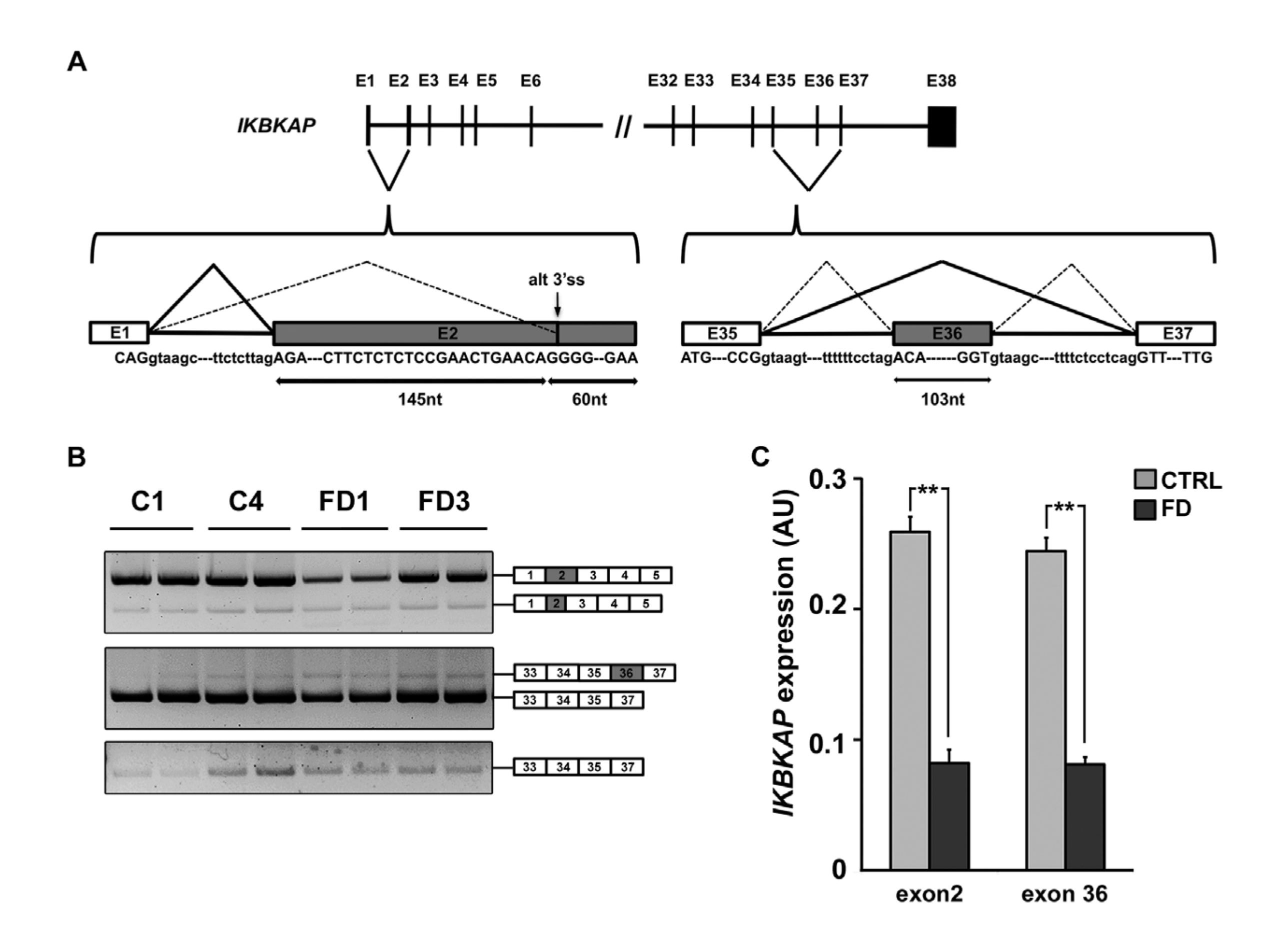


Figure 6. Expression and alternative splicing of *IKBKAP* mRNA at the extremities of the coding sequence. Two additional splicing events are described within *IKBKAP* gene. The first one represents the alternative use of a 3'ss for exon 2 and the second one concerns exon 36 skipping, as

represented by a schematic (A). B, semi-quantitative RT-PCR illustrates relative amounts of both events on 2 control and 2 FD hOE-MSCs. C, RT-qPCR analysis performed at 2 different regions (exon 2 and exon 36), using 4 control and 4 FD samples, at the same cell passage (P7). Histograms represent the mean value of 4 samples, normalized with *ABL1* gene. (*** $P<0{,}001$ using two-tailed Student's test). IKAP/hELP1 truncated regions for all splicing events are represented by a schematic (D). Grey portions represent the conserved amino acids while black portions represent new amino acids resulting from a frame shift. Putative functional domains of the protein are indicated as well as the immunogenic region for the monoclonal antibody used in western blot and immunocytochemistry experiments.
doi:10.1371/journal.pone.0015590.g006

However, quantitative analysis by RT-qPCR revealed that kinetin significantly enhances the ratio 6 h after its addition to the culture (Figure 7E). Interestingly, time-response of kinetin was maximal at 24 h during treatment, but its effect on splicing lasted more than 6 h after the drug was washed out and WT transcript levels remained high compared to non-treated cells, at least 24 h after the wash out. Consistent results were observed for IKAP/hELP1 protein expression by western blot analysis, although a strong decrease of protein amount appeared after 24 h of wash out (Figure 7F).

Finally, we wanted to investigate kinetin activity along the *IKBKAP* transcript. Therefore, we compared the level of expression of *IKBKAP* transcripts by RT-qPCR to focus on different transcript regions for both control and FD hOE-MSCs, with or without the presence of 80 μM kinetin for 24h. We first observed that the total amount of *IKBKAP* transcript detected was almost identical when probes at the extremities or in the middle of the transcript were used (Figure 7G). In addition, kinetin has no significant effects on *IKBKAP* transcript levels in control cells, which likely excludes a potential action of kinetin on *IKBKAP* transcription. Moreover, kinetin, by improving *IKBKAP* exon 20 recognition, restores *IKBKAP* transcript levels in FD hOE-MSCs similar to those observed in control cells (Figure 7G). Kinetin did not modify the ratio of alternative splicing around exon 2 and exon 36, suggesting its specific mechanism of action on exon 20 inclusion (data not shown).

Altogether, these results revealed that kinetin exerts a rapid and possibly long lasting effect on *IKBKAP* mRNA splicing, which most likely occurs by increasing *IKBKAP* mRNA stability rather than acting on transcription.

FD sphere cells display a strongly reduced *IKBKAP* exon 20 skipping

One property of multipotent cell consists in their capacity to organize into spheres when cultured in appropriate medium. Since FD hOE-MSCs express a significant amount of MU *IKBKAP* transcript we asked whether induction of sphere formation could modify the WT:MU *IKBKAP* transcripts ratio.

Although hOE-MSCs proliferate as adherent cells, when cultured in DMEM-F12 supplemented with serum (Figure 8A), they progressively organize into spherical aggregates when cultured in serum-free medium in the presence of EGF and bFGF (Figure 8B and C). Both control and FD hOE-MSCs were able to form spheres in approximately one week, and immunostaining with anti-β-III tubulin (Figure 8D) and anti-nestin (Figure 8E) antibodies revealed a similar staining of both markers for control and FD cells (Figure 8F). Total RNAs, isolated from either FD spheres, FD cells cultured in serum during the same period, or dissociated cells from spheres that were reintroduced in serum medium for 24 h, were subjected to RT-qPCR. We observed a significant increase of *IKBKAP* exon 20 inclusion in spheres, when compared to hOE-MSCs in serum conditions, as well as a semi-disappearance of *IKBKAP* exon 20 skipping (Figure 8G). Dissociated spheres re-exposed to serum rapidly expressed initial levels of WT and MU transcripts (Figure 8G). We quantified WT and MU transcript level of expression in these 3 different conditions and confirmed that spheres formation from FD hOE-MSC induces *IKBKAP* mRNA splicing correction, using RT-qPCR (Figure 8H). We also looked for exon 2 and exon 36 alternative splicing events but did not detect significant alterations of splicing ratio resulting from sphere formation and dissociation (data not shown).

Commitment of FD OE-MSCs into neuronal and glial lineages leads to a more severe *IKBKAP* exon 20 skipping

FD hOE-MSCs were treated for 7 days to induce neuronal differentiation with a protocol previously used in hOE-MSC [41], which consists of additing retinoic acid, forskolin, and Sonic hedgehog in the medium (called rafnshh medium). Cells were first cultured in serum-free medium supplemented with N2 and B27 until they became adherent before being cultured in rafnshh (Figure 9B). The new culture medium induced a slight morphology change, as compared to the serum condition (Figure 9A). When hOE-MSCs were first cultured in rafnshh, they began to form long fine processes and neural-like cells (Figure 9C). After 7 days of treatment, a majority of cells adopted neuron-like morphologies (Figure 9D) and established a wide range of connections (Figure 9E, F and M). Using end-point PCR on 3 different FD cell cultures, we observed that *IKBKAP* mRNA splicing in rafnshh-treated cells was more prone to exon 20 skipping as compared to untreated cells (Figure 9G). This change can be quantified by RT-qPCR (Figure 9H). In contrast, we did not detect significant variations in exon 2 and exon 36 alternative splicing during neuronal differentiation (data not shown).

When assessing immunostaining on treated cells, we observed that rafnshh treatment increased the proportion of both GFAP (Figure 9I and J) and MAP2-expressing cells (Figure 9K and L). Double-labeling with β-III tubulin and nestin revealed a stronger expression of β-III tubulin compared to nestin during the differentiation process (Figure 9N–P). Collectively, these results indicate that rafnshh treatment influences the neural and glial lineage commitment. As a consequence, the splicing machinery in neuron or astrocyte-differentiated cells is impaired for *IKBKAP* exon 20 recognition.

Discussion

Deciphering the molecular basis of the tissue-specific pattern of *IKBKAP* mRNA splicing in FD nervous tissues is crucial for the comprehension of disease physiopathology in this genetic neurological disorder affecting neuronal development and survival. In this study, we aimed to recapitulate different aspects of *IKBKAP* gene expression using FD hOE-MSCs cultured with different conditions. While other human cellular models, such as fibroblasts or iPS cells, have been investigated to understand FD, we believe that hOE-MSCs hold a great promise to model the FD disease pathology. hOE-MSCs are easily obtained by a simple biopsy and can be maintained for an extended period of time and can be rapidly expanded in basic culture conditions without genetic manipulation. In addition, due to the origins of hOE-MSCs from a peripheral tissue, these cells are able to express neuroglial markers *in vitro* (Figure 1, [42–47]). Thus, they constitute an efficient and simple method to derive neuronal cells in the original context of the genetic mutation studied.

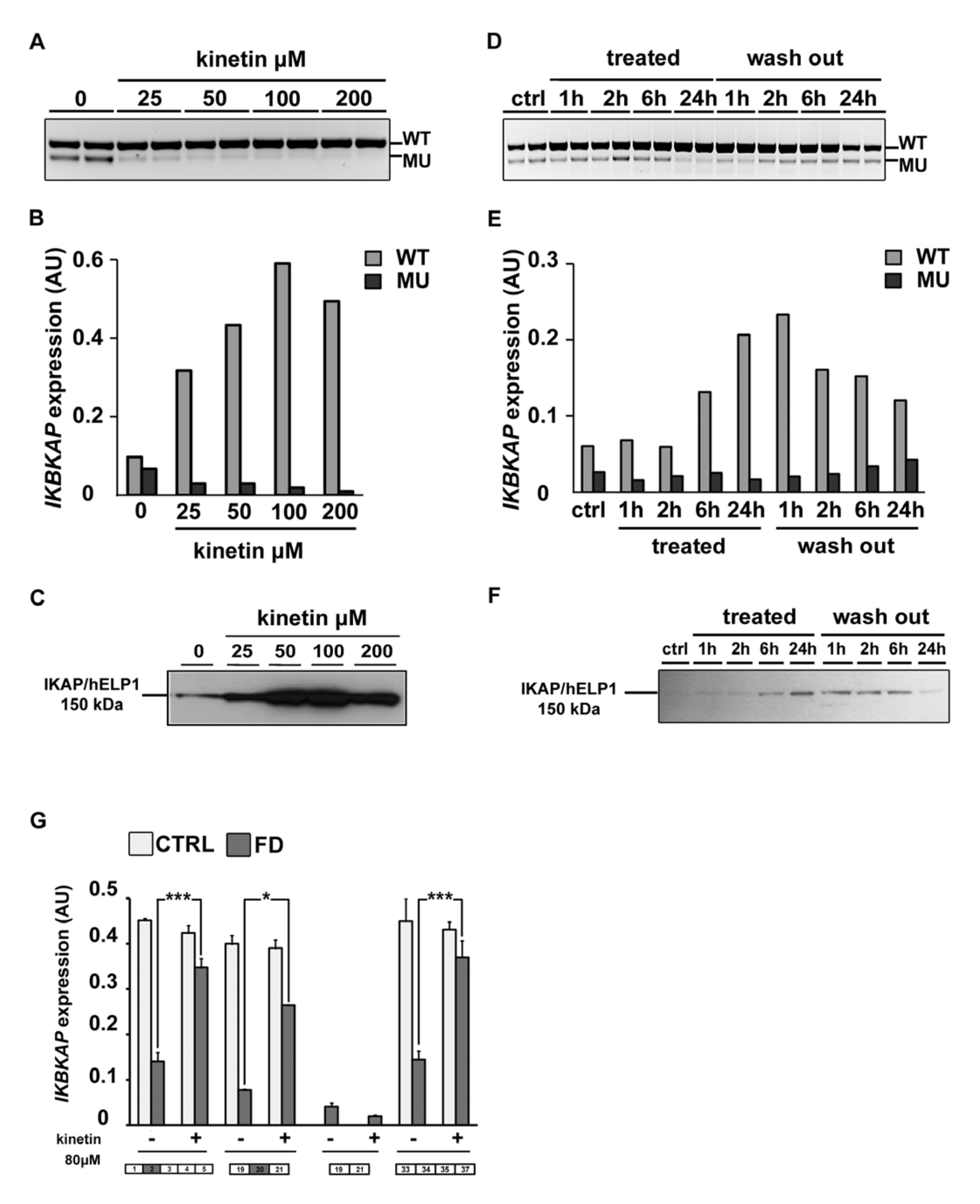


Figure 7. Action of kinetin on *IKBKAP* mRNA splicing in FD hOE-MSCs cells. A–C, hOE-MSCs were treated with increasing concentration of kinetin for 72 h. (i) Total RNAs were reverse transcribed and subjected to both semi-quantitative PCR (A) and absolute qPCR (B) of WT and MU *IKBKAP* transcripts, (ii) total lysates were analyzed by western blot using a monoclonal mouse anti-IKAP/hELP1 antibody (C). D–F, kinetics of hOE-MSCs incubated for 24 h with 80 µM kinetin which was then removed for the next 24 h. Total RNAs were reverse-transcribed and subjected to *IKBKAP*-specific semi-quantitative PCR (D) and absolute RT-qPCR (E). Total lysates were analyzed by western blot (F). The level of WT and MU transcripts was normalized using *ABL1* as a reference gene (B and E). G, Two controls (CTRL) and two FD hOE-MSCs treated or not with 80 µM kinetin for 24h were analyzed by absolute RT-qPCR to determine the amount of *IKBKAP* exon 2 inclusion, exon 20 inclusion, exon 20 skipping, and exon 36 exclusion after normalization with *ABL1*. (* $P<0.05$; *** $P<0,001$ using two-tailed Student's test).
doi:10.1371/journal.pone.0015590.g007

In contrast to control cells, which constitutively include *IKBKAP* exon 20, we confirmed the alternative splicing of that exon in FD cells (Figure 2A). Similar to neural precursors obtained from iPS cells [22], FD hOE-MSCs predominantly express the MU *IKBKAP* transcript isoform (Figure 2A and B). Moreover, we demonstrated that FD cells exhibit notably lower *IKBKAP* transcript levels (WT+MU), when compared to controls (Figures 2B, 6C and 7G). Such a difference is most likely explained by an extensive degradation of MU transcripts through the NMD pathway (Figure 2D), as was previously suggested [35]. However, NMD efficiency varies between cell types and individuals [48–50] and it is unclear how prevalent this mechanism of mRNA degradation is in the nervous system of FD individuals. In order to get a better insight into the actual contribution of NMD on the decay of *IKBKAP* MU transcripts, it will be necessary to specifically block the NMD machinery.

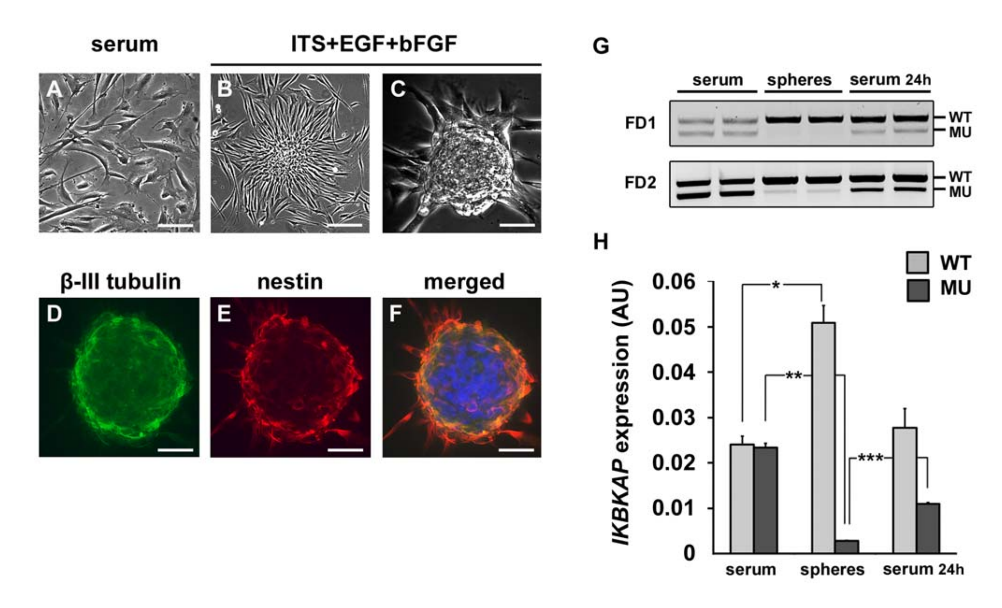


Figure 8. WT:MU ratio is increased in hOE-MSC-derived spheres. FD hOE-MSCs cultured in serum (A) gave rise to spheres when plated in medium supplemented with ITS and growth factors (B and C). Immunostaining showed β-III tubulin and nestin positive spheres (D, E, F). RNA was isolated from 2 different FD hOE-MSCs cultured first in serum, then induced to form spheres, and finally dissociated and replated in serum conditions for 24h (G). RT-qPCR was performed on the same samples and histograms represent the mean value of the two FD samples after normalization with *ABL1* gene (H). Scale bars represent 100 μm. (* $P<0{,}05$, ** $P<0{,}01$ using two-tailed Student's test).
doi:10.1371/journal.pone.0015590.g008

So far, the proposed functions of IKAP/hELP1 are related to various cellular localizations. This has been the matter of a controversy because several studies have failed to detect IKAP/hELP1 in the nucleus or found it almost exclusively in the cytoplasm [11,21,34,51], which is difficult to reconcile with its suggested role in transcription elongation. As observed in most published studies, we observed that the immunolocalization of IKAP/hELP1 was mainly cytoplasmic within the perinuclear area. However, we also detected significant nuclear staining in agreement with other reported studies [9,52]. Altogether, our findings of IKAP/hELP1 distribution in hOE-MSCs support multiple roles for the protein within different subcellular compartments.

In order to establish a direct link between low levels of *IKBKAP* WT transcripts and decreased neuronal populations in FD patients, several groups have investigated transcriptome variations resulting from a reduced level of *IKBKAP* transcripts using microarray technology [10,22,34,40,53]. However, poor correlations were observed between these studies. Several reasons can explain these discrepancies. First, various cell types, used at different stages of development and differentiation, have been studied (brain tissue, fibroblasts, HeLa cells, HCT116 cells, iPS cells). The cells tested in the current study are likely at a stage between the iPS and iPS cell-derived neural crest precursors developed by the Studer group [22]. It is thus not surprising that most of our microarray results overlap with that of the iPS cell study. Second, a potential source of variability among transcriptome analyses derives from the technical manipulations employed to downregulate *IKBKAP* (presence of the FD mutation in its original context compared to WT *IKBKAP* knockdown using different interfering RNAs), resulting in differential residual *IKBKAP*/IKAP expression. Third, in the context of a rare disease, a small sample size may cast doubt upon the validity of drawn conclusions. To decrease statistical bias, we decided to increase the number of samples of our 5 control and 4 FD patients by collecting data from 4 different passages (P1, P2, P5 and P9) of each primary cell line. We hypothesized was that such a method would allow us to i) increase the statistical power of our analysis, and ii) explore the effects of culture conditions on gene expression. We anticipated that a strong gene dysregulation observed in microarray would be more significant if this expression is stably maintained at any cell passage. Most of the differentially expressed genes were found to have a modest (<2) fold-change (Table 1). Interestingly, like previous studies, we observed that a majority of genes were down-regulated in FD hOE-MSCs (Table 1, negative values), and only 4 genes were up-regulated (Table 1 and Table S1, positive values). This observation is in agreement with other studies [10,22,40] and may reflect a defect in transcription due to decreased Elongator activity, as previously proposed [10]. Importantly, 10 genes in our list (20%) appeared to be correlated with one or two previous investigations (Table 1 and Table S1, last column). In one of the past studies, *IKBKAP* expression level can be downregulated by RNAi in control cells [10], where there is no production of MU transcripts. Thus, different studies share dysregulated genes in different contexts of either constitutive or alternative splicing of *IKBKAP* mRNA. This suggests that *IKBKAP* alternative splicing may not be the only pathological alteration in FD.

Similar to what was previously reported, our study revealed the downregulation of gelsolin (*GSN*), a protein involved in cell motility, that causes defects in cytoskeleton reorganization and cell migration in FD [10,11,22,34]. The most dysregulated gene observed in our study was *PMEPA1* (−4.92 fold change), encoding the TMEPAI protein which has recently been reported to be a direct target of the TGF-β signaling pathway and is involved in cell growth, cell differentiation, and apoptosis [54]. Due to its important cellular function and repeated reports of its dysregulation in FD cells, it would be very interesting to test TMEPAI in further studies. In agreement with previous studies correlating a decreased expression of IKAP/hELP1 with defects in cell migration [10,11,22,34], the Boyden's chamber assay show that FD hOE-MSCs have decreased migration potential, compared to control cells (Figure 5).

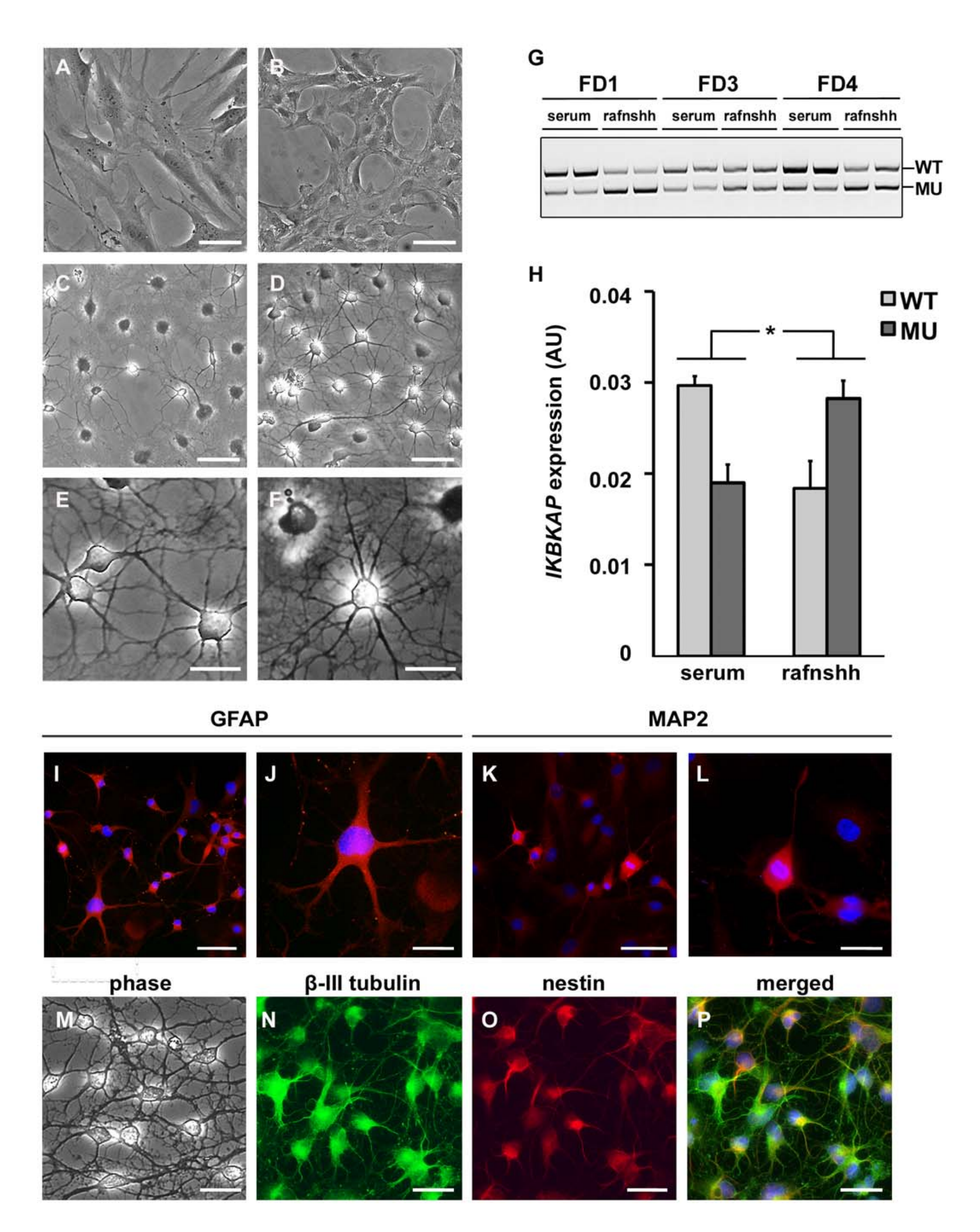


Figure 9. WT:MU ratio is decreased in differentiated hOE-MSCs. A, Phase contrast microscopy of FD hOE-MSCs cultured for 48 h in either serum (A), or N2B27 conditions (B), or with the rafnshh cocktail including retinoic acid, forskolin and Sonic hedgehog (C), or rafnshh for 7 days (D). Details of connections established between cells and extensive cellular arborization after 7 days in rafnshh condition are shown in E and F. G, agarose gel electrophoresis of semi-quantitative RT-PCR products showing *IKBKAP* WT and MU transcripts of 3 differents FD patients before and after rafnshh treatment. H, histograms represent the mean level of *IKBKAP* transcripts expression normalized with *ABL1* gene expression for 3 FD patients, after RT-qPCR analysis. (* $P<0,05$, using two-tailed Student's test). FD hOE-MSCs treated for 7 days in rafnshh were fixed and stained for GFAP (I, one cell with ramified neuritic process is magnified in J), MAP2 expression (K and L). M–P, double labelling of rafnshh-treated cells with anti-β III tubulin and anti-nestin antibodies. Scale bars represent either 50 µm (A–D, I, K, M–P) or 25 µm (E–F, J, L).
doi:10.1371/journal.pone.0015590.g009

Surprisingly, we did not find *IKBKAP* as a dysregulated gene in our microarray analysis. This result is all the more intriguing since this gene is expressed at much lower levels in FD hOE-MSCs, as shown by RT-qPCR in the exon 20 region (Figure 2B). However, previous analyses using microarrays also failed to detect *IKBKAP* as a down-regulated gene in FD cells [22,40]. To address the question of a possible PCR artifact or lowered microarray sensitivity and because the FD mutation is located in the middle region of the *IKBKAP* gene, we performed quantitative PCR at both ends of the *IKBKAP* gene.

Unexpectedly, we identified 2 new events of alternative splicing at both extremities of *IKBKAP* coding sequence (Figure 6A). However, these results are in agreement with EST sequences found in alternative splicing database such as ASD [55]. We

revealed that the alternative use of a 3′ss (downstream of the ATG start codon) leads to a shorter exon 2 which can potentially induce the use of an alternative ATG start codon (in exon 4), resulting in the synthesis of an N-terminal truncated IKAP/hELP1 protein. In addition, we detected the presence of intronic sequences at the end of *IKBKAP* gene leading to a supplementary exon in the mRNA, named "exon 36". This exon inclusion also induced a frameshift and resulted in a premature stop codon whose relative location likely led to NMD of this new isoform, as observed by stabilization of the transcript after cycloheximide treatment (Figure S2).

IKAP/hELP1 plays the role of a scaffold protein in Elongator complex assembly and the C-terminus half of IKAP/hELP1 is responsible for this function [34]. It has also been shown that IKAP/hELP1 contains five WD-like repeats domains in the N-terminal part that may play a role for protein-protein interactions [56]. When comparing the different protein isoforms resulting from the 3 alternative splicing events we described (Figure 6D), only the isoform resulting from exon 20 skipping seems to lack a functional domain and may play a pathological role during FD progression. However, the protein domains of IKAP/hELP1 important for Elongator integrity have not been precisely mapped and other roles for IKAP/hELP1 have been proposed outside of the Elongator complex. Thus, it is not clear whether the new *IKBKAP* isoforms we described, may have functional roles. Future investigations with specific reagents (antibodies) will be required to address this issue.

Nevertheless, we consistently detected a lower expression of *IKBKAP* gene (including the full length exon 2 transcript and the transcripts skipping exon 36) in FD hOE-MSCs (Figure 6C), as determined when investigating the exon 20 region. Thus, the relatively stable expression of *IKBKAP* observed in microarray analysis may be due to a weak expression that is masked within the noise signals. Furthermore, during the analysis of our microarray data and those of previous studies [10,22,40] we noticed that a high fraction of genes were expressed at background levels. This points to the limitation of using microarray technology to establish the whole genome expression pattern. We expect that new technologies such as RNA deep sequencing will rival PCR sensitivity and specificy in the near future.

The model of hOE-MSCs from FD patients has also been very useful to test compounds, such as kinetin, that can correct the defective splicing process. As reported in the other cell types tested, we confirmed that kinetin corrects splicing in a dose-dependent manner in FD hOE-MSCs (Figure 7A–C). This suggests that kinetin activity is not cell-type specific. Although the mechanism by which kinetin modulate splicing remains poorly understood [36,57], it is unlikely that kinetin acts directly on the general transcription machinery as the level of *IKBKAP* transcripts was not significantly modulated by kinetin in control hOE-MSCs (Figure 7G). This effect of kinetin has also been previously observed in control iPS cells [22]. Time-course experiments of kinetin treatment revealed that the drug acts quite rapidly on correcting *IKBKAP* mRNA splicing and enhancing IKAP/hELP1 synthesis, but its effects last only a short time after removal (Figure 7E and F). This information provides new perspectives in the strategy of kinetin delivery to FD patients. First, kinetin as an FD treatment would potentially decrease deleterious consequences of the mutation at the protein level. In addition, drug efficacy may be achieved if adequate levels of kinetin is maintained over a long period of time. However, as observed for FD iPS cells [22], kinetin did not improve cell migration in FD hOE-MSCs (Figure 5), suggesting incomplete phenotype complementation.

Using the hOE-MSCs model, we were also able to modulate the expression of *IKBKAP* WT and MU transcripts, by exposing the cells to different culture conditions to simulate variations in alternative splicing occurring during development and differentiation. hOE-MSCs form free floating spheres in approximately 7 days, when cultured with EGF and bFGF in serum-free medium. It is known that in sphere conditions, cells can form a niche prevent differentiation and ensure self-renewal. The cell populations contained in hOE-MSC-derived spheres are not well-known. Some reports indicate that they include an heterogeneous mixture of stem cells and neuroglial progenitors [29,43,44,47,58]. However, immunostainings of nestin and β-III tubulin show no significant differences. Interestingly, PCR analysis demonstrated that spheres express higher levels of WT *IKBKAP* transcript compared to hOE-MSCs in serum and express very flow levels of MU transcript. However, when the cells were transferred back to culture conditions with serum, the enhanced *IKBKAP* exon 20 inclusion was not maintained (Figure 8G and H). FD hOE-MSCs that were cultured in serum-free conditions without forming spheres did not exhibit significant changes in *IKBKAP* isoforms, suggesting that there is subpopulation of cells within the spheres that can promote *IKBKAP* exon 20 inclusion. These results indicate that when FD cells are turned back into a more "primitive" developmental stage, *IKBKAP* aberrant splicing is corrected, as was described during the fibroblast to iPS cell reprogramming process [22]. Accordingly, commitment into a more differentiated neuronal state would alter *IKBKAP* exon 20 inclusion. Therefore, we differentiated FD hOE-MSCs, using a previously established protocol which included retinoic acid (RA), forskolin (FN), and Sonic hedgehog (Shh) in the culture medium [44]. In these conditions, we observed that differentiated cells express the highest levels of MU *IKBKAP* transcript (Figure 9G and H). This result correlates with the specific low WT:MU *IKBKAP* isoform ratio in nervous tissues [8], and suggests that stem cells engaged in a neuronal lineage with appropriate culture conditions can rapidly switch their *IKBKAP* WT:MU transcript ratio.

Previous studies have shown that i) *IKBKAP* exon 20 is poorly defined in a "healthy" context, due to the presence of a weak 3′ss and exonic splicing silencers, and ii) the FD mutation exacerbates the environment leading to alternative exon 20 inclusion in FD tissues [59,60]. We propose that some transcription/splicing factors involved in *IKBKAP* exon 20 recognition are also downregulated in a tissue-specific manner. This would explain why the pattern of *IKBKAP* alternative mRNA splicing is more aberrant in the nervous system. Interestingly, Lee and colleagues determined that the neuron-specific splicing factor NOVA1 [61] was underexpressed in FD versus control iPS cell-derived neural crest precursors [22]. The new model described in this study will allow us further test whether candidate splicing factors may be involved in the tissue-specific regulation of *IKBKAP* mRNA alternative splicing.

Materials and Methods

Ethics Statement

All control and FD participants gave informed and written consents (provided by the parents for the children) and biopsies were obtained under a protocol, which was approved by the local ethical committees in New York (Institutional Review Board of the New York University School of Medicine) and Marseille (Comité Consultatif de Protection des Personnes dans la Recherche Biomédicale Marseille 2).

Purification of hOE-MSCs

Human nasal mucosae were obtained from biopsies of 4 FD patients (3 females and 1 male aged 12–16 years) at the

Dysautonomia Treatment and Evaluation Center, New York. All four FD patients were homozygous for the splicing mutation. Biopsies form 5 healthy controls (3 females and 2 males, aged 18–39 years) were collected by the ENT Department in Marseille (Hopital Nord, France). Biopsies were harvested as previously described [30,46] to obtain a cell culture of hOE-MSCs. The cells were continuously cultured in DMEM/HAM'S F12 (Gibco) supplemented with 10% fetal bovine serum (FBS) and 50 μg/ml gentamicin (Gibco) and trypsinized once a week with 0.05% trypsin-EDTA (Gibco) at 60–80% confluence. Cycloheximide (Sigma), diluted in DMSO, was used at 50 μg/ml. Kinetin solution (Sigma, 1 mg/ml) was diluted in DMEM/HAM'S F12 at concentration ranging from 25 to 200 μM for various incubation times, as specified in the text.

Generation of spheres and cell differentiation

Cells were plated at 15,000 cells/cm^2 into 6-well plates preteated with poly-L-lysine (5 μg/cm^2, Sigma) in a serum-free medium of DMEM supplemented with insulin-transferrin-selenium (ITS, 1g/l insulin, 0.55 g/l transferrin, 0.67 mg/l sodium selenite; Gibco), epidermal growth factor (EGF, 50 ng/ml, R&D system) and basic fibroblast growth factor-2 (bFGF, 50 ng/ml, R&D system). Half of the medium was changed every 2 days. Multipotent spheres were obtained after 1 week and harvested by aspiration of the culture medium and centrifugation (5 min, 300g). They were then incubated in Accumax solution (Sigma), for 10 min at 37°C. To release more cells, the sample was gently triturated by repeated pipetting. When disaggregation was complete, cells were centrifugated (5 min, 300g) to remove cell debris. For cell differentiation, hOE-MSCs were plated on glass coverslips at the density of 10,000 cells/cm^2 (in six-well plates for RNA extraction, and 24-well plates for immunostaining), in serum-free medium supplemented with 1% ITS, 1% B27 and 0.5% N2, until adhesion. Cells were then treated with 1% ITS, 1 μM *all-trans* retinoic acid (Sigma), 5 μM Forskolin (R&D Systems), 15 nM Sonic hedgehog (R&D Systems), 1% B27 and 0.5% N2 for 7 days without changing the medium.

Immunocytochemistry

Cells grown on glass coverslips were fixed with 4% paraformaldehyde for 20 min at room temperature and rinsed three times with phosphate-buffered saline (PBS). Cells were preincubated for 60 min at room temperature with blocking buffer (3% BSA in PBS with 0.1% Triton X-100 and 10% normal goat serum), followed by incubation with the primary antibodies diluted in the blocking buffer. Coverslips were processed for immunofluorescence staining using the following primary antibodies: rabbit anti-nestin (1:500, Abcys), mouse anti βIII-tubulin (1:500, Sigma, clone SDL.3D10), rabbit anti-GFAP (1:500, Dako), rabbit anti-MAP2 (1:500, Abcam), mouse anti-IKAP/hELP1 (1:100, BD Biosciences, clone 33). Each primary antibody was applied for 2 h at room temperature. For IKAP/hELP1 staining, primary antibody was incubated 3 h at room temperature followed by an overnight incubation at 4°C. We used appropriate secondary antibodies: goat anti-rabbit IgG conjugated with AlexaFluor 594 (1:500, Invitrogen), goat anti-mouse IgG conjugated with AlexaFluor 488 (1:500, Invitrogen) for 1 h at room temperature. Hoechst nuclear dye was used to label nuclei (1:2,000, Molecular Probes, #33258). Coverslips were finally mounted with anti-fading medium (ProLong®, invitrogen). Cells were observed under a Nikon Eclipse E800 upright microscope equipped with epifluorescence and TRITC, FITC and DAPI filters, and images were analyzed using an Orca-ER CCD camera (Hamamatsu Photonics) and the LUCIA image analysis software (Laboratory Imaging). Confocal image acquisition was performed on a Leica TCS SP2 confocal microscope (Leica Microsystems) using the 488-nm band of an argon laser for excitation of Alexa 488 and the 680-nm band of an argon laser for excitation of Alexa 680. High magnification images were acquired using a 63× HCX PL APO (with 4 digital zoom factor) oil immersion objective (numerical aperture 1.32) by sequential scanning to minimize the crosstalk of fluorophores. Pinhole size was set to 'Airy one'' to achieve the best possible resolution (theoretical lateral and axial limits: 165 and 330 nm, respectively). Voxel size was set to 58 nm in x and y and to 162 nm in z.

Western blot analysis

Cells were harvested by trypsination and centrifugation (5 min, 300g). The pellet, containing approximately 10^6 cells, was resuspended in 0.5 ml 2× Laemmli buffer (0.5 M Tris pH 6.8, 4.4 ml Glycerol, 20% SDS, 1% Bromophenol Blue, 0.5 ml β-mercaptoethanol). 30 μl of cell lysates were separated on 6.5% SDS-polyacrylamide gel electrophoresis and transferred to a nitrocellulose membrane (Amersham Biosciences). After blocking with 5% nonfat milk in PBS, 0.1%, tween 20 (PBST) buffer, blots were probed for 1h at room temperature with a mouse monoclonal anti-IKAP antibody (1:5,000, BD Biosciences, clone 33) in PBST, followed by incubation with horseraddish peroxidase-conjugated goat anti-mouse IgG (1:5,000, Jackson Immunoresearch) for 45 min at room temperature. As a control, the membrane was also probed for β-actin (1:3,000, Sigma, clone SDL.3010). Proteins were visualized by chemiluminescent detection using the ECL detection kit (Enhanced Chemiluminescence, Amersham) and films were digitized and analyzed using the Bio 1D software.

RNA Isolation and semi-quantitative reverse transcription-polymerase chain reaction analysis

Total RNA was isolated using the RNeasy Mini Kit (Qiagen) with DNAse treatment on the column according to manufacturer's recommendation. Total RNA was subjected to reverse transcription (RT) using the High-Capacity cDNA Archive Kit (Applied Biosystems). End-point polymerase chain reaction (PCR) analysis was performed using the Go-Taq polymerase system (Promega) and *IKBKAP*-specific primers listed in Table 2. PCR products were separated on a 1.7% agarose gel by electrophoresis in 1× TBE buffer (Tris 0.89 M, boric acid 0.89 M and EDTA 0,02 M). DNA was visualized under UV light after ethidium bromide incorporation and documented using BioVision Camera.

Plasmid calibrators

A fragment of WT *IKBKAP* cDNA, containing exon 19-exon 20-exon 21 and the 16 first nt of intron 21, was cloned into pcDNA 3.1 TOPO vector (Invitrogen) and named *IKBKAP* cDNA cal. Similarly, a piece of MU *IKBKAP* cDNA, containing exon 19-exon 21 and the 19 first nt of exon 22, was cloned into a pcDNA 3.1 TOPO vector and named *IKBKAP* skipEx20cal. A piece of WT *IKBKAP* cDNA, containing the last 103 nt of exon 35-exon37-exon38 first 90 nt, was cloned into KpnI-XbaI cloning sites of pcDNA 3.1 TOPO vector and named *IKBKAP* skipEx36cal. A piece of WT *IKBKAP* cDNA, containing the last 30 nt of exon1-exon2-exon3 first 110 nt, was cloned into KpnI-XbaI cloning sites of pcDNA 3.1 TOPO vector and named *IKBKAP* fullEx2cal. For *ABL1*, the last 37 nt of exon 2 and first 102 nt of exon 3 were amplified from *ABL1* cDNA, cloned into KpnI-XbaI cloning sites of pcDNA 3.1 TOPO vector and named *ABL1* cal. All plasmid calibrators were linearized with XbaI and serially diluted in a

Table 2. Sequence of primers used for end-point and TaqMan real-time PCR.

Primer Probe	Sequence	T (°C)[a]	Amplicon size (bp)	Splicing events
End-point PCR				
hIKBKAP ex17-18F	TCATCAATGACATTGAGGTTG	55	446 (WT)	ex20 incl/skip
hIKBKAP ex22R	ATGATTCACAGAATCTATCTG		372 (MU)	
hIKBKAP ex1F	CCGGACGCACCTCTGTTTG	60	485	alt 3′ss for ex2
hIKBKAP ex4-5R	TCAGGGTCTGTTGACCTGTG		340 (alt 3′ss ex2)	
hIKBKAP ex33-34F	TCCAGGATATCAGCGAGATC	59	449	ex36 incl/skip
hIKBKAP ex37R	GCTGATAAGATGCCATGATAC		346 (- ex36)	
hIKBKAP ex35-36R	TTGGGACCTAGAACACCTGT	59	414	ex36 incl
Real-time PCR				
hELP1 ex19F	GGTTCACGGATTGTCACTGTT	60	133	ex20 incl
hELP1 ex20-21R	ACATAAGTTTGTCCAACCACTTCC			
P-WTELP1 ex20R	AAACCAGGGCTCGATGATGAACA			
hELP1 ex19-21F	GGACACAAAGCTTGTATTACAGACTTA	60	121	ex20 skip
hELP1 ex21-22R	CCACATTTCCAAGAAACACCT			
P-MUELP1 ex21F	AGAGGCATTTGAATGCATGAGAAAGC			
hELP1 ex2F	CCAGGGAATCCTCAGTGCT	60	104	full length ex2 incl
hELP1 ex2-3R	TTCACTTCTCTTGAGACAGGGTCTAC			
P-WTELP1 ex2F	TCCGACTGAACAGGGGACGGT			
hELP1 ex35-37F	CAGCTACCCCGGTTCTAGGT	60	128	ex36 skip
hELP1 ex38R	GGTTCTTCTGTTGATCTTTGGTG			
P-WTELP1 ex37-38R	AAGCTCAGCATCAAGAACAGGAACC			

[a]Annealing temperature.

doi:10.1371/journal.pone.0015590.t002

solution of Tris 10 mM, EDTA 1 mM pH 8, containing 20 ng/µl of *E. coli* 16S and 23S rRNA (Roche).

Real-time PCR assay

The PCR reactions were performed in triplicate in a final volume of 25 µl, including 300 nM primers, 200 nM TaqMan® probe, 12.5 µl of TaqMan® universal PCR master mix (Applied Biosystems) and 5 µl of either cDNA or plasmid calibrator in a AB Prism 7900 HT thermocycler with 50 cycles and the protocol recommended by the manufacturer. For relative quantification and microarray results validation, we selected primer sets and probes, matching sequences present in the IMAGE human cDNA clones of the nylon microarrays with those displayed on the web portal of Applied Biosystems. The assay IDs were the following: Hs00375306_m1 (*PMEPA1*) and Hs00293488_m1 (*S100A16*) for the dysregulated genes in FD, and Hs01003267_m1 (*HPRT1*) and Hs00293488_m1 (*RPLP0*) for reference genes used to normalize the data. We also used previously validated primers and probe for *ABL1* as a third reference gene [62]. Results were calculated using the $2(-\Delta\Delta C_T)$ method [63]. For absolute quantification, *IKBKAP* primers and hydrolysis probes (FAM TAMRA) were designed using the Primer 3 software and are listed in Table 2. Serial dilutions of plasmid calibrators (10^6, 10^5, 10^4, 10^3, 10^2 copies in 5 µl) were prepared and used to construct the standard curves. The number of *IKBKAP* and *ABL1* transcripts was extrapolated automatically by the Sequence Detection System v2.2.2 software (Applied Biosystems).

Microarray analysis and normalization

RNA integrity was assessed using an Agilent 2100 Bioanalyser (Palo Alto, CA). Samples with an RNA integrity number (RIN)$<$9 were excluded from the analysis (the samples concerned were C2P5, C3P5, and FD2P5). Gene expression analyses were carried out with cDNA Nylon microarrays containing 8,780 spotted cDNA clones and radioactive detection as previously described [64], with 5 µg of RNA reverse transcribed (oligo-dT priming) in presence of [α-^{33}P]dCTP (Amersham Pharmacia Biotech). Details about microarray construction, clones list, probes preparations, hybridizations and washes have been previously described [65]. After image acquisition, signal intensities were quantified using BZScan software (http://tagc.univ-mrs.fr/bioinformatics/bzscan, [66]). A specific R library that uses the 'S4' system of formal classes and methods was used to process and normalize nylon microarray data [67]. Quantile normalization was applied to vector probe data (V) and complex probe data (C), to correct for global intensity and dispersion. Correction by the vector signal was made for each spot signal by calculating a C:V ratio before log transformation (base 2). No background correction or overshining correction was used. All data are MIAME compliant and have been loaded into ArrayExpress database (http://www.ebi.ac.uk/microarray-as/ae/) under accession number E-MTAB-281.

Statistical and gene ontology analysis

Significant Analysis of Microarray (SAM version 1.13; Standford University) was applied to determine significant differential gene expression using the Multiexperiment viewer (MEV) program. The data were analyzed using a two-class unpaired response type, which compared control samples versus FD samples. SAM calculated a significant score for each gene based on the gene expression change relative to the standard deviation of repeated values for that gene. We used 100 permutations and a

false discovery rate (FDR) of 3%. A total of 50 genes appearing in the heat map generation were called as significant with a p-value<0.006. For gene ontology analysis, we generated a set of human protein associated with the gene appearing as significant with the SAM test, by using the BioKnowledge® Library (BKL) Retriever™ search tool (http://www.biobase-international.com/). This set of proteins was analyzed for overrepresentation of Gene Ontology (GO) Biological Process (BP) terms.

Boyden chamber-based cell migration assay

hOE-MSCs were detached by trypsin/EDTA, counted and seeded into the upper chamber of transwell polyethylene terephtalate filter membranes with 8 µM diameter pores (BD Biosciences), at a density of 3×10^4 cells/well, in a final volume of 200 µl serum or serum free culture medium, with or without 100 µM kinetin. Cells were allowed to migrate through the membrane filter for 24h at 37°C, 5% CO_2. Cells migrating thought the membrane pore and invading the underside surface of the membrane were fixed with 4% paraformaldehyde. Non-migratory cells on the upper membrane surface were removed with a cotton swab, and nuclei were stained with 0.5 µg/mL DNA intercalant Hoechst #33258. For quantitative assessment, the number of stained migrating cells was counted with imageJ software on 10 random fields per membrane filter at ×20 magnification.

Supporting Information

Figure S1 50 genes are differentially expressed between control and FD hOE-MSCs. Heatmap representation of overexpressed (red) and underexpressed (green) genes in 5 controls and 4 FD OE-MSCs at passage 1,2,5, and 9. Normalized signal intensities were treated with the SAM software to highlight the most differentially expressed genes, with a FDR set at 3%.
(TIF)

Figure S2 *IKBKAP* exon 36 inclusion increases after cycloheximide treatment. NMD pathway was blocked by the translation inhibitor cycloheximide and results in an elevated expression of exon 36-including transcripts in 2 FD OE-MSC cultures (FD3 and FD4) as determined by absolute RT-qPCR. (* $P<0.05$).
(TIF)

Table S1 Dysregulated genes involved in other processes.
(DOC)

Acknowledgments

We wish to thank the patients and their families for their contribution to this study. Furthermore, we thank Dr. Joseph Bernstein, Dr. Arnaud Deveze and Dr. Jacques Magnan for their support in collecting biopsies. We also thank Denis Puthier for his expertise in bioinformatics analysis, as well as Adlane Ould-Yahoui for technical help, André Verdel and Jeanne Hsu for critical reading of the manuscript.

Author Contributions

Conceived and designed the experiments: NB FF FBA ECI. Performed the experiments: NB BL AB OS CFT ECI. Analyzed the data: NB ECI. Contributed reagents/materials/analysis tools: JG MK CN FF FBA. Wrote the paper: NB ECI.

References

1. Axelrod FB, Iyer K, Fish I, Pearson J, Sein ME, et al. (1981) Progressive sensory loss in familial dysautonomia. Pediatrics 67: 517–22.
2. Pearson J, Pytel BA, Grover-Johnson N, Axelrod F, Dancis J (1978) Quantitative studies of dorsal root ganglia and neuropathologic observations on spinal cords in familial dysautonomia. J Neurol Sci 35: 77–92.
3. Pearson J, Pytel BA (1978) Quantitative studies of sympathetic ganglia and spinal cord intermedio-lateral gray columns in familial dysautonomia. J Neurol Sci 39: 47–59.
4. Axelrod FB (2004) Familial dysautonomia. Muscle Nerve 29: 352–63.
5. Anderson SL, Coli R, Daly IW, Kichula EA, Rork MJ, et al. (2001) Familial dysautonomia is caused by mutations of the IKAP gene. Am J Hum Genet 68: 753–8.
6. Slaugenhaupt SA, Blumenfeld A, Gill SP, Leyne M, Mull J, et al. (2001) Tissue-specific expression of a splicing mutation in the IKBKAP gene causes familial dysautonomia. Am J Hum Genet 68: 598–605.
7. Dong J, Edelmann L, Bajwa AM, Kornreich R, Desnick RJ (2002) Familial dysautonomia: detection of the IKBKAP IVS20(+6T→C) and R696P mutations and frequencies among Ashkenazi Jews. Am J Med Genet 110: 253–7.
8. Cuajungco MP, Leyne M, Mull J, Gill SP, Lu W, et al. (2003) Tissue-specific reduction in splicing efficiency of IKBKAP due to the major mutation associated with familial dysautonomia. Am J Hum Genet 72: 749–58.
9. Hawkes NA, Otero G, Winkler GS, Marshall N, Dahmus ME, et al. (2002) Purification and characterization of the human elongator complex. J Biol Chem 277: 3047–52.
10. Close P, Hawkes N, Cornez I, Creppe C, Lambert CA, et al. (2006) Transcription impairment and cell migration defects in elongator-depleted cells: implication for familial dysautonomia. Mol Cell 22: 521–31.
11. Creppe C, Malinouskaya L, Volvert ML, Gillard M, Close P, et al. (2009) Elongator controls the migration and differentiation of cortical neurons through acetylation of alpha-tubulin. Cell 136: 551–64.
12. Solinger JA, Paolinelli R, Kloss H, Scorza FB, Marchesi S, et al. (2010) The Caenorhabditis elegans Elongator complex regulates neuronal alpha-tubulin acetylation. PLoS Genet 6: e1000820.
13. Chen C, Tuck S, Bystrom AS (2009) Defects in tRNA modification associated with neurological and developmental dysfunctions in Caenorhabditis elegans elongator mutants. PLoS Genet 5: e1000561.
14. Huang B, Johansson MJ, Bystrom AS (2005) An early step in wobble uridine tRNA modification requires the Elongator complex. RNA 11: 424–36.
15. Esberg A, Huang B, Johansson MJ, Bystrom AS (2006) Elevated levels of two tRNA species bypass the requirement for elongator complex in transcription and exocytosis. Mol Cell 24: 139–48.
16. Rahl PB, Chen CZ, Collins RN (2005) Elp1p, the yeast homolog of the FD disease syndrome protein, negatively regulates exocytosis independently of transcriptional elongation. Mol Cell 17: 841–53.
17. Okada Y, Yamagata K, Hong K, Wakayama T, Zhang Y (2010) A role for the elongator complex in zygotic paternal genome demethylation. Nature 463: 554–8.
18. Lipardi C, Paterson BM (2009) Identification of an RNA-dependent RNA polymerase in Drosophila involved in RNAi and transposon suppression. Proc Natl Acad Sci U S A 106: 15645–50.
19. Chen YT, Hims MM, Shetty RS, Mull J, Liu L, et al. (2009) Loss of mouse Ikbkap, a subunit of elongator, leads to transcriptional deficits and embryonic lethality that can be rescued by human IKBKAP. Mol Cell Biol 29: 736–44.
20. Hims MM, Shetty RS, Pickel J, Mull J, Leyne M, et al. (2007) A humanized IKBKAP transgenic mouse models a tissue-specific human splicing defect. Genomics 90: 389–96.
21. Valensi-Kurtz M, Lefler S, Cohen MA, Aharonowiz M, Cohen-Kupiec R, et al. (2010) Enriched population of PNS neurons derived from human embryonic stem cells as a platform for studying peripheral neuropathies. PLoS One 5: e9290.
22. Lee G, Papapetrou EP, Kim H, Chambers SM, Tomishima MJ, et al. (2009) Modelling pathogenesis and treatment of familial dysautonomia using patient-specific iPSCs. Nature 461: 402–6.
23. Saha K, Jaenisch R (2009) Technical challenges in using human induced pluripotent stem cells to model disease. Cell Stem Cell 5: 584–95.
24. Vierbuchen T, Ostermeier A, Pang ZP, Kokubu Y, Sudhof TC, et al. (2010) Direct conversion of fibroblasts to functional neurons by defined factors. Nature 463: 1035–41.
25. Kim K, Doi A, Wen B, Ng K, Zhao R, et al. (2010) Epigenetic memory in induced pluripotent stem cells. Nature 467: 285–90.
26. Polo JM, Liu S, Figueroa ME, Kulalert W, Eminli S, et al. (2010) Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. Nat Biotechnol 28: 848–55.
27. Ghosh Z, Wilson KD, Wu Y, Hu S, Quertermous T, et al. (2010) Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. PLoS One 5: e8975.

28. Graziadei PP, Graziadei GA (1979) Neurogenesis and neuron regeneration in the olfactory system of mammals. I. Morphological aspects of differentiation and structural organization of the olfactory sensory neurons. J Neurocytol 8: 1–18.
29. Murrell W, Feron F, Wetzig A, Cameron N, Splatt K, et al. (2005) Multipotent stem cells from adult olfactory mucosa. Dev Dyn 233: 496–515.
30. Delorme B, Nivet E, Gaillard J, Haupl T, Ringe J, et al. (2010) The human nose harbors a niche of olfactory ectomesenchymal stem cells displaying neurogenic and osteogenic properties. Stem Cells Dev 19: 853–66.
31. Feron F, Perry C, Hirning MH, McGrath J, Mackay-Sim A (1999) Altered adhesion, proliferation and death in neural cultures from adults with schizophrenia. Schizophr Res 40: 211–8.
32. McCurdy RD, Feron F, Perry C, Chant DC, McLean D, et al. (2006) Cell cycle alterations in biopsied olfactory neuroepithelium in schizophrenia and bipolar I disorder using cell culture and gene expression analyses. Schizophr Res 82: 163–73.
33. Murrell W, Wetzig A, Donnellan M, Feron F, Burne T, et al. (2008) Olfactory mucosa is a potential source for autologous stem cell therapy for Parkinson's disease. Stem Cells 26: 2183–92.
34. Johansen LD, Naumanen T, Knudsen A, Westerlund N, Gromova I, et al. (2008) IKAP localizes to membrane ruffles with filamin A and regulates actin cytoskeleton organization and cell migration. J Cell Sci 121: 854–64.
35. Slaugenhaupt SA, Mull J, Leyne M, Cuajungco MP, Gill SP, et al. (2004) Rescue of a human mRNA splicing defect by the plant cytokinin kinetin. Hum Mol Genet 13: 429–36.
36. Hims MM, Ibrahim EC, Leyne M, Mull J, Liu L, et al. (2007) Therapeutic potential and mechanism of kinetin as a treatment for the human splicing disease familial dysautonomia. J Mol Med 85: 149–61.
37. Gold-von Simson G, Goldberg JD, Rolnitzky LM, Mull J, Leyne M, et al. (2009) Kinetin in familial dysautonomia carriers: implications for a new therapeutic strategy targeting mRNA splicing. Pediatr Res 65: 341–6.
38. Tanabe S, Sato Y, Suzuki T, Suzuki K, Nagao T, et al. (2008) Gene expression profiling of human mesenchymal stem cells for identification of novel markers in early- and late-stage cell culture. J Biochem 144: 399–408.
39. Wagner W, Horn P, Castoldi M, Diehlmann A, Bork S, et al. (2008) Replicative senescence of mesenchymal stem cells: a continuous and organized process. PLoS One 3: e2213.
40. Cheishvili D, Maayan C, Smith Y, Ast G, Razin A (2007) IKAP/hELP1 deficiency in the cerebrum of familial dysautonomia patients results in down regulation of genes involved in oligodendrocyte differentiation and in myelination. Hum Mol Genet 16: 2097–104.
41. Zhang X, Klueber KM, Guo Z, Cai J, Lu C, et al. (2006) Induction of neuronal differentiation of adult human olfactory neuroepithelial-derived progenitors. Brain Res 1073–1074: 109–19.
42. Wolozin B, Sunderland T, Zheng BB, Resau J, Dufy B, et al. (1992) Continuous culture of neuronal cells from adult human olfactory epithelium. J Mol Neurosci 3: 137–46.
43. Roisen FJ, Klueber KM, Lu CL, Hatcher LM, Dozier A, et al. (2001) Adult human olfactory stem cells. Brain Res 890: 11–22.
44. Zhang X, Klueber KM, Guo Z, Lu C, Roisen FJ (2004) Adult human olfactory neural progenitors cultured in defined medium. Exp Neurol 186: 112–23.
45. Winstead W, Marshall CT, Lu CL, Klueber KM, Roisen FJ (2005) Endoscopic biopsy of human olfactory epithelium as a source of progenitor cells. Am J Rhinol 19: 83–90.
46. Feron F, Perry C, McGrath JJ, Mackay-Sim A (1998) New techniques for biopsy and culture of human olfactory epithelial neurons. Arch Otolaryngol Head Neck Surg 124: 861–6.
47. Othman M, Lu C, Klueber K, Winstead W, Roisen F (2005) Clonal analysis of adult human olfactory neurosphere forming cells. Biotech Histochem 80: 189–200.
48. Viegas MH, Gehring NH, Breit S, Hentze MW, Kulozik AE (2007) The abundance of RNPS1, a protein component of the exon junction complex, can determine the variability in efficiency of the Nonsense Mediated Decay pathway. Nucleic Acids Res 35: 4542–51.
49. Bateman JF, Freddi S, Nattrass G, Savarirayan R (2003) Tissue-specific RNA surveillance? Nonsense-mediated mRNA decay causes collagen X haploinsufficiency in Schmid metaphyseal chondrodysplasia cartilage. Hum Mol Genet 12: 217–25.
50. Resta N, Susca FC, Di Giacomo MC, Stella A, Bukvic N, et al. (2006) A homozygous frameshift mutation in the ESCO2 gene: evidence of intertissue and interindividual variation in Nmd efficiency. J Cell Physiol 209: 67–73.
51. Holmberg C, Katz S, Lerdrup M, Herdegen T, Jaattela M, et al. (2002) A novel specific role for I kappa B kinase complex-associated protein in cytosolic stress signaling. J Biol Chem 277: 31918–28.
52. Kim JH, Lane WS, Reinberg D (2002) Human Elongator facilitates RNA polymerase II transcription through chromatin. Proc Natl Acad Sci U S A 99: 1241–6.
53. Cornez I, Creppe C, Gillard M, Hennuy B, Chapelle JP, et al. (2008) Deregulated expression of pro-survival and pro-apoptotic p53-dependent genes upon Elongator deficiency in colon cancer cells. Biochem Pharmacol 75: 2122–34.
54. Watanabe Y, Itoh S, Goto T, Ohnishi E, Inamitsu M, et al. (2010) TMEPAI, a transmembrane TGF-beta-inducible protein, sequesters Smad proteins from active participation in TGF-beta signaling. Mol Cell 37: 123–34.
55. Stamm S, Riethoven JJ, Le Texier V, Gopalakrishnan C, Kumanduri V, et al. (2006) ASD: a bioinformatics resource on alternative splicing. Nucleic Acids Res 34: D46–55.
56. Cohen L, Henzel WJ, Baeuerle PA (1998) IKAP is a scaffold protein of the IkappaB kinase complex. Nature 395: 292–6.
57. Pros E, Fernandez-Rodriguez J, Benito L, Ravella A, Capella G, et al. (2009) Modulation of aberrant NF1 pre-mRNA splicing by kinetin treatment. Eur J Hum Genet 18: 614–7.
58. Viktorov IV, Savchenko EA, Chekhonin VP (2007) Spontaneous neural differentiation of stem cells in culture of human olfactory epithelium. Bull Exp Biol Med 144: 596–601.
59. Carmel I, Tal S, Vig I, Ast G (2004) Comparative analysis detects dependencies among the 5′ splice-site positions. RNA 10: 828–40.
60. Ibrahim EC, Hims MM, Shomron N, Burge CB, Slaugenhaupt SA, et al. (2007) Weak definition of IKBKAP exon 20 leads to aberrant splicing in familial dysautonomia. Hum Mutat 28: 41–53.
61. Ule J, Ule A, Spencer J, Williams A, Hu JS, et al. (2005) Nova regulates brain-specific splicing to shape the synapse. Nat Genet 37: 844–52.
62. Beillard E, Pallisgaard N, van der Velden VH, Bi W, Dee R, et al. (2003) Evaluation of candidate control genes for diagnosis and residual disease detection in leukemic patients using 'real-time' quantitative reverse-transcriptase polymerase chain reaction (RQ-PCR) - a Europe against cancer program. Leukemia 17: 2474–86.
63. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(−Delta Delta C(T)) Method. Methods 25: 402–8.
64. Ballester B, Ramuz O, Gisselbrecht C, Doucet G, Loi L, et al. (2006) Gene expression profiling identifies molecular subgroups among nodal peripheral T-cell lymphomas. Oncogene 25: 1560–70.
65. Talby L, Chambost H, Roubaud MC, N'Guyen C, Milili M, et al. (2006) The chemosensitivity to therapy of childhood early B acute lymphoblastic leukemia could be determined by the combined expression of CD34, SPI-B and BCR genes. Leuk Res 30: 665–76.
66. Lopez F, Rougemont J, Loriod B, Bourgeois A, Loi L, et al. (2004) Feature extraction and signal processing for nylon DNA microarrays. BMC Genomics 5: 38.
67. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.

RESEARCH ARTICLE

Human Mutation

OFFICIAL JOURNAL



Genome-Wide Analysis of Familial Dysautonomia and Kinetin Target Genes with Patient Olfactory Ecto-Mesenchymal Stem Cells

Nathalie Boone,[1†] Aurélie Bergon,[2†] Béatrice Loriod,[2] Arnaud Devèze,[3] Catherine Nguyen,[2] Felicia B Axelrod,[4] and El Chérif Ibrahim[1*]

[1]*Aix-Marseille Université, NICN, UMR 6184, Marseille, France, CNRS, NICN, UMR 6184, Marseille, France;* [2]*TAGC, INSERM UMR_S 928, Aix-Marseille Université, Marseille, France;* [3]*Département ORL, Hôpital Universitaire Nord, AP-HM, Marseille, France;* [4]*Department of Pediatrics, New York University School of Medicine, New York, NY*

Communicated by Mireille Claustres

Received 10 October 2011; accepted revised manuscript 8 December 2011.
Published online 20 December 2011 in Wiley Online Library (www.wiley.com/humanmutation).DOI: 10.1002/humu.22010

ABSTRACT: Familial dysautonomia (FD) is a rare inherited neurodegenerative disorder. The most common mutation is a c.2204+6T>C transition in the 5′ splice site (5′ss) of *IKBKAP* intron 20, which causes a tissue-specific skipping of exon 20, resulting in lower synthesis of IKAP/hELP1 protein. To better understand the specificity of neuron loss in FD, we modeled the molecular mechanisms of *IKBKAP* mRNA splicing by studying human olfactory ecto-mesenchymal stem cells (hOE-MSCs) derived from FD patient nasal biopsies. We explored how the modulation of *IKBKAP* mRNA alternative splicing impacts the transcriptome at the genome-wide level. We found that the FD transcriptional signature was highly associated with biological functions related to the development of the nervous system. In addition, we identified target genes of kinetin, a plant cytokinin that corrects *IKBKAP* mRNA splicing and increases the expression of IKAP/hELP1. We identified this compound as a putative regulator of splicing factors and added new evidence for a sequence-specific correction of splicing. In conclusion, hOE-MSCs isolated from FD patients represent a promising avenue for modeling the altered genetic expression of FD, demonstrating a methodology that can be applied to a host of other genetic disorders to test the therapeutic potential of candidate molecules.
Hum Mutat 00:1–11, 2012. © 2011 Wiley Periodicals, Inc.

KEY WORDS: familial dysautonomia; *IKBKAP*; RNA splicing; transcriptome analysis

Introduction

Familial dysautonomia (FD, Riley-Day syndrome, hereditary sensory and autonomic neuropathy type III, MIM# 223900) is a rare neurodegenerative disease with autosomal recessive inheritance and a carrier frequency of 1 in 31 in the Ashkenazi Jewish population [Scott et al., 2010]. The disease is characterized by anatomical selective depletion of sensory and autonomic neurons [Axelrod et al., 1981; Pearson and Pytel, 1978; Pearson et al., 1978] resulting in variable symptoms including: decreased sensitivity to pain, lack of overflow tearing, inappropriate blood pressure control manifested as orthostatic hypotension and episodic hypertension, poor oral coordination resulting in poor feeding and swallowing, and gastrointestinal dysmotility [Axelrod, 2004]. FD is a disease for which no cure is currently available, and treatment is aimed at controlling symptoms and prevention of complications.

FD is caused by mutations in the *IKBKAP* gene (MIM# 603722), which encodes a protein termed IKAP/hELP1 [Anderson et al., 2001; Slaugenhaupt et al., 2001]. The most prevalent mutation, is the T-to-C transition in position six of the 5′ splice site (5′ss) of intron 20 (c.2204+6T>C), occurring in >99.5% of cases of FD [Anderson et al., 2001; Dong et al., 2002; Scott et al., 2010; Slaugenhaupt et al., 2001]. This mutation leads to a tissue-specific skipping of exon 20 of *IKBKAP* mRNA (MU isoforms). The defective splicing leads to low levels of transcripts including exon 20 (WT isoforms), reduced synthesis of IKAP/hELP1 protein, and appears to be more severe in sensory and autonomic nervous systems than others tissues [Cuajungco et al., 2003].

IKAP/hELP1 was identified as the scaffold protein required to assemble a well-conserved six-protein complex (ELP1-6), also called the holo-Elongator complex [Hawkes et al., 2002], which is recruited to the transcribed regions of some human genes essentially involved in actin cytoskeleton regulation and cell motility migration. Subsequently, IKAP/hElongator was also shown to have functions in cell migration [Close et al., 2006; Creppe et al., 2009], acetylation of microtubules, and neuronal development [Solinger et al., 2010]. It was also proposed to play a role in exocytosis [Rahl et al., 2005], and zygotic paternal genome demethylation [Okada et al., 2010], but most likely as a result of tRNA modifications [Chen et al., 2009a; Esberg et al., 2006; Huang et al., 2005; Li et al., 2009].

Several studies aimed at investigating transcriptional alterations revealed distinct patterns of gene expression in FD. Indeed, a subgroup of genes associated with cell migration and actin cytoskeleton was shown to be downregulated in IKAP/hElp1 deficient HeLa and

Additional Supporting Information may be found in the online version of this article.

†These authors contributed equally to this work.

*Correspondence to: El Chérif Ibrahim, 51 Bd Pierre Dramard, 13344 Marseille Cedex 15 France. E-mail: el-cherif.ibrahim@univmed.fr

Contract grant sponsor: Association Française de Recherche contre les Myopathies (AFM).

FD fibroblast cells [Close et al., 2006]. Others identified genes known to be involved in oligodendrocyte development, myelin formation, and disorganization of microtubules from cerebrum of FD patients [Cheishvili et al., 2007, 2011]. Lee and colleagues determined that the neuron-specific splicing factor *NOVA1* was underexpressed in FD versus control-induced pluripotent stem cell (iPSC) derived neural crest precursors [Lee et al., 2009]. Finally, a recent study showed that FD affects genes important for early developmental stages of the nervous system using neuroblastoma cell lines [Cohen-Kupiec et al., 2011]. Nevertheless, the specific means by which aberrant *IKBKAP* mRNA splicing causes the disease producing developmental and degenerative neuronal changes in FD neurons is still unclear. However, the plant cytokinin kinetin has been found to be a powerful agent that corrects *IKBKAP* mRNA splicing defects [Boone et al., 2010; Hims et al., 2007; Keren et al., 2010; Lee et al., 2009; Slaugenhaupt et al., 2004] and was effective when administered in transgenic mouse model [Shetty et al., 2011] and FD patients [Axelrod et al., 2011], which would make it a potential therapeutic agent for the treatment of FD and other disorders involving missplicing of mRNAs.

To better understand the cascade of events mediated by the c.2204+6T>C mutation, we used human olfactory ectomesenchymal stem cells (hOE-MSCs) from FD patients or from control individuals as an experimental model. This allowed us to modulate the rate of *IKBKAP* exon 20 skipping in vitro by varying culture conditions to produce spheres (with epidermal growth factor (EGF), and basic fibroblast growth factor (bFGF)) or to stimulate neuroglial differentiation (with a "rafnshh" cocktail including *all-trans* retinoic acid, forskolin, and sonic hedgehog) [Boone et al., 2010]. In this study, we performed the comparative transcriptome analysis between spheres and rafnshh-treated hOE-MSCs and also investigated the effect of kinetin at the genome-wide level.

Materials & Methods

Purification of hOE-MSCs

Human nasal mucosae were obtained by biopsying five FD patients (four females and one male aged 10–16 years) at the Dysautonomia Treatment and Evaluation Center, New York. Biopsies from five healthy controls (four females and one male, aged 10–34 years) were collected by the ENT Department in Marseille (University Hôpital Nord, France). Samples were obtained under a protocol approved by the local ethical committees in New York and Marseille. Biopsies were harvested as previously described [Boone et al., 2010] to obtain an olfactory cell culture of hOE-MSCs. Cells are routinely cultivated with DMEM/HAM'S F12 containing 10% FBS at 37°C in the presence of 5% CO2. Kinetin solution (1 mg/ml, Sigma-Aldrich, St. Louis, MO) was diluted in DMEM/HAM'S F12 at 100 μM concentration for dose-effect experiments, and at 80 μM in experiments of consecutive addition and washout of kinetin. For transcriptome analysis, four of the five control and FD hOE-MSCs have been used.

Generation of Spheres and Induction of Cell Differentiation

Multipotent spheres were obtained after 1 week of culture with EGF and bFGF as previously described [Boone et al., 2010]. For cell differentiation, hOE-MSCs were treated with the rafnshh cocktail consisting in 1% insulin-transferrin-selenium (ITS), 1 μM *all-trans* retinoic acid (Sigma-Aldrich), 5 μM Forskolin (R&D Systems, Minneapolis, MN), 15 nM Sonic hedgehog (R&D Systems), 1% B27 supplement (a serum substitute), and 0.5% N2 supplement (enhancing the growth and survival of neuronal cells) for 7 days without changing the medium.

RNA Isolation

Total RNA was isolated using the RNeasy Mini Kit (Qiagen, Hilden, Germany) with DNAse treatment on the column following the manufacturer's guidelines. RNA concentration was determined using a nanodrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE). RNA integrity was assessed on an Agilent 2100 Bioanalyzer (Palo Alto, CA). All samples exhibited RIN>9.

End-Point Reverse Transcription-Polymerase Chain Reaction Analysis

Total RNA was subjected to reverse transcription (RT) using the High-Capacity cDNA Archive Kit (Applied Biosystems, Foster City, CA). End-point polymerase chain reaction (PCR) analysis was performed using the Go-Taq® DNA polymerase system (Promega, Madison, WI) and *IKBKAP*-specific primers (hIKBKAP 17-18F and hIKBKAP 22R; see Boone et al. [2010]). PCR products were separated on a 1.7% agarose gel by electrophoresis in 1X TBE buffer (Tris 0.89 M, boric acid 0.89 M, and EDTA 0.02 M). DNA was visualized under UV light after ethidium bromide incorporation and documented using BioVision Camera.

Real-Time PCR Assay

The PCR reactions were performed in duplicate in a final volume of 25 μl, including 300-nM primers, 200-nM TaqMan® probe, 12.5 μl of TaqMan® universal PCR master mix (Applied Biosystems) and 25–50 ng of cDNA in a AB Prism 7900 HT thermocycler with 50 cycles and the protocol recommended by the manufacturer. Primers hELP1 ex19F, hELP1 ex20-21R, and probe P-WTELP1 ex20R were used for detection of *IKBKAP* transcripts containing exon 20, while primers hELP1 ex19-21F, hELP1 ex21-22R, and probe P-MUELP1 ex21F were used for detection of *IKBKAP* transcripts skipping exon 20 [Boone et al., 2010]. To determine the level of expression of candidate genes dysregulated genes in FD, the following primer/TaqMan probe assays were obtained from Applied Biosystems: Hs_00176719m1 (*LYN*), Hs_01103338m1 (*SNCA*), Hs_01374916m1 (*MAP1LC3C*), Hs_00359592m1 (*NOVA1*), Hs_01120488m1 (*SPON1*), Hs_00216077m1 (*LUC7L*), Hs_00214302m1 (*ZNF280D*), and Hs00296608_m1 (*WDR59*) was used as a reference gene to normalize the data. Results were calculated using the $2(-\Delta\Delta C_T)$ method [Livak and Schmittgen, 2001].

Preparation of Samples and Microarray Assay

Sample amplification, labeling, and hybridization essentially followed the one-color microarray-based gene expression analysis (low input quick amp labeling) protocol (version 6.5, May 2010) recommended by Agilent Technologies. In brief, 500 ng of each total RNA sample was reverse transcribed into cDNA using oligo dT-T7 promoter primer. Labeled cRNA was synthesized from the cDNA. The reaction was performed in a solution containing dNTP mix, cyanine 3-dCTP, and T7 RNA Polymerase, and incubated at 40°C for 2 hr. Hybridization was performed into whole human genome microarray slides (4 × 44K G4112F, Agilent Technologies, Santa Clara, CA) containing 45,220 oligonucleotide probes at 65°C for 17 hr. Hybridized microarray slides were then washed according to the

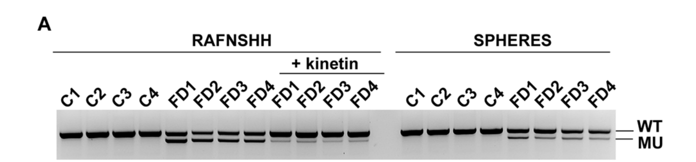
manufacturer's instructions and scanned using an Agilent DNA Microarray Scanner, using the Agilent Feature Extraction Software (Agilent Technologies). The microarray data are available from the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) under the series accession number GSE27915.

Microarray Data Analysis

Quantification files derived from the Agilent Feature Extraction Software were analyzed using the AgiND package (http://tagc.univ-mrs.fr/AgiND). We also used the AgiND R package for quality control and normalization. Quantile methods and a background correction were used for data normalization.

Statistical Analysis

For each comparison ("spheres" vs. "rafnshh", "controls" vs. "FD", "control spheres" vs. "FD spheres", and "FD rafnshh" vs. "FD rafnshh treated with kinetin"), measurement of differential gene expression was obtained using the Multiexperiment viewer (MEV) program. Significant Analysis of Microarray (SAM version 1.13; Standford University) and Student's *t*-test were applied to determine fold changes (FC) and *P*-values (*P*), respectively. The data were analyzed using a two-class unpaired response type, which compared control versus FD samples as well as untreated versus treated FD samples with kinetin. To construct dendrograms, average linkage approximate hierarchical clustering of genes was performed using Pearson correlation (using Cluster [Eisen et al., 1998]) and visualized under Treeview software (http://jtreeview.sourceforge.net/). For each comparison of samples, the statistically relevant signaling pathways, corresponding to the differentially expressed genes, were identified using DAVID (Database for Annotation, Visualization, and Integrated Discovery; http://david.abcc.ncifcrf.gov/) [Huang et al., 2009] with high classification stringency, $P < 0.05$ and FDR < 20%.



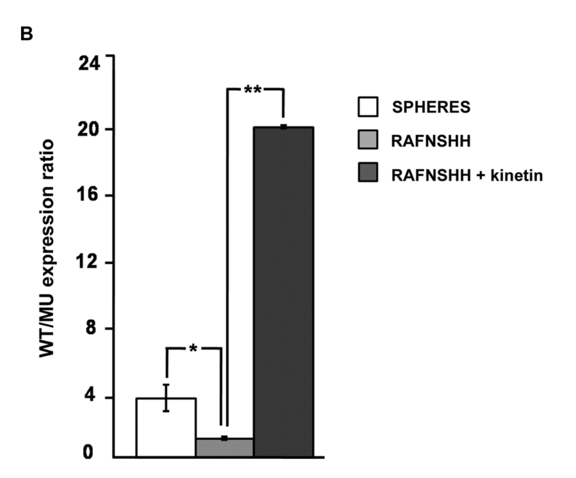


Figure 1. Expression profile of *IKBKAP* exon 20 alternative splicing in control and FD hOE-MSCs under defined culture conditions. **A:** Agarose gel electrophoresis of semi quantitative RT-PCR products obtained from four control and four FD hOE-MSCs cultivated in sphere and kinetin treated/untreated differentiation (rafnshh) conditions. *IKBKAP* transcripts are identified as WT for the correct transcript and MU for the exon 20-skipped transcript. **B:** Relative RT-qPCR was performed using cDNAs from the same samples of the three conditions ($P < 0.01$, * $P < 0.05$). Ct mean values for all samples from each condition were used and normalized with Ct mean values of *WDR59*.

Results

IKBKAP Splice Variants Ratio is Affected by Culture Conditions and Kinetin in FD hOE-MSCs

To observe the variation in *IKBKAP* mRNA alternative splicing, four control and four FD hOE-MSC cultures were either induced to form spheres or treated with the rafnshh cocktail. FD rafnshh-treated hOE-MSCs were also incubated with 100 μM kinetin for 48 hr. A semi-quantitative RT-PCR analysis confirmed that control hOE-MSCs expressed exclusively the WT *IKBKAP* mRNA transcript while FD hOE-MSCs expressed both the WT and the MU transcripts (Fig. 1A). In contrast, RT-qPCR analysis on the FD samples revealed a reduced WT/MU transcript expression ratio in rafnshh compared to sphere conditions, which was reversed with kinetin treatment (Fig. 1B). These results are consistent with the increased WT *IKBKAP* transcripts observed in spheres compared to adherent hOE-MSCs from our previous study [Boone et al., 2010].

Microarray Analysis Revealed Differential Transcriptional Expression of *IKBKAP* and Genes Implicated in Nervous System Function

The 16 RNA samples obtained after treating four control and four FD hOE-MSCs with either EGF and bFGF, or the rafnshh cocktail, were used to characterize the FD transcriptional signature. To confirm the strong impact of culture conditions on gene expression, we compared the control and FD sphere samples to the rafnshh-treated samples (without kinetin). After conducting a significant analysis of microarray (SAM) analysis, we visualized as a heatmap that more than 3,000 transcripts are differentially expressed (false discovery rate, FDR = 0) between spheres and neuroglial progenitors (Supp. Fig. S1). Of these genes, we analyzed only those with a more than 10-fold change (FC) superior and grouped them under five types of biological processes: nervous system development, cell adhesion, WNT/Shh signaling pathway, proteolysis, and retinoic acid activity (Supp. Table S1). All of the processes appear to be related to the factors added in culture media. Indeed, genes that show the greatest fold changes are involved in retinoic acid activity (*RARRES1, DHRS3, RARRES2, RARB)* and the WNT/Shh signaling pathway (*SFRP4, CP, WNT11*). In general, genes related to the nervous system are more highly expressed in spheres in comparison to the differentiated samples. In addition, many genes involved in proteolysis were upregulated in spheres samples (*MMP1*, *MMP10*, *ADAMST14*, *MME*, *PRSS35*, and *ADAMST8*).

Using the SAM analysis, we next compared the FD signature between control and FD samples. We assumed a FDR of 10% and characterized 35 differentially expressed genes with a FC>2 (Fig. 2). Although most of the genes were downregulated in FD, *IKBKAP* appears the second most discriminant marker between control and FD hOE-MSCs. Importantly, 10 differentially expressed genes encode proteins playing important role in neural cells: *CD40*, *FXYD1*, *GPR37*, *LYN*, *NRG1*, *PACSIN1*, *RUNX3*, *SCN2B*, *SFRP2*, *SNCA* [Aubert et al., 2002; Burré et al., 2010; Deng et al., 2007; Gibb et al., 2011; Hossain et al., 2010; Kramer et al., 2006; Lopez-Santiago

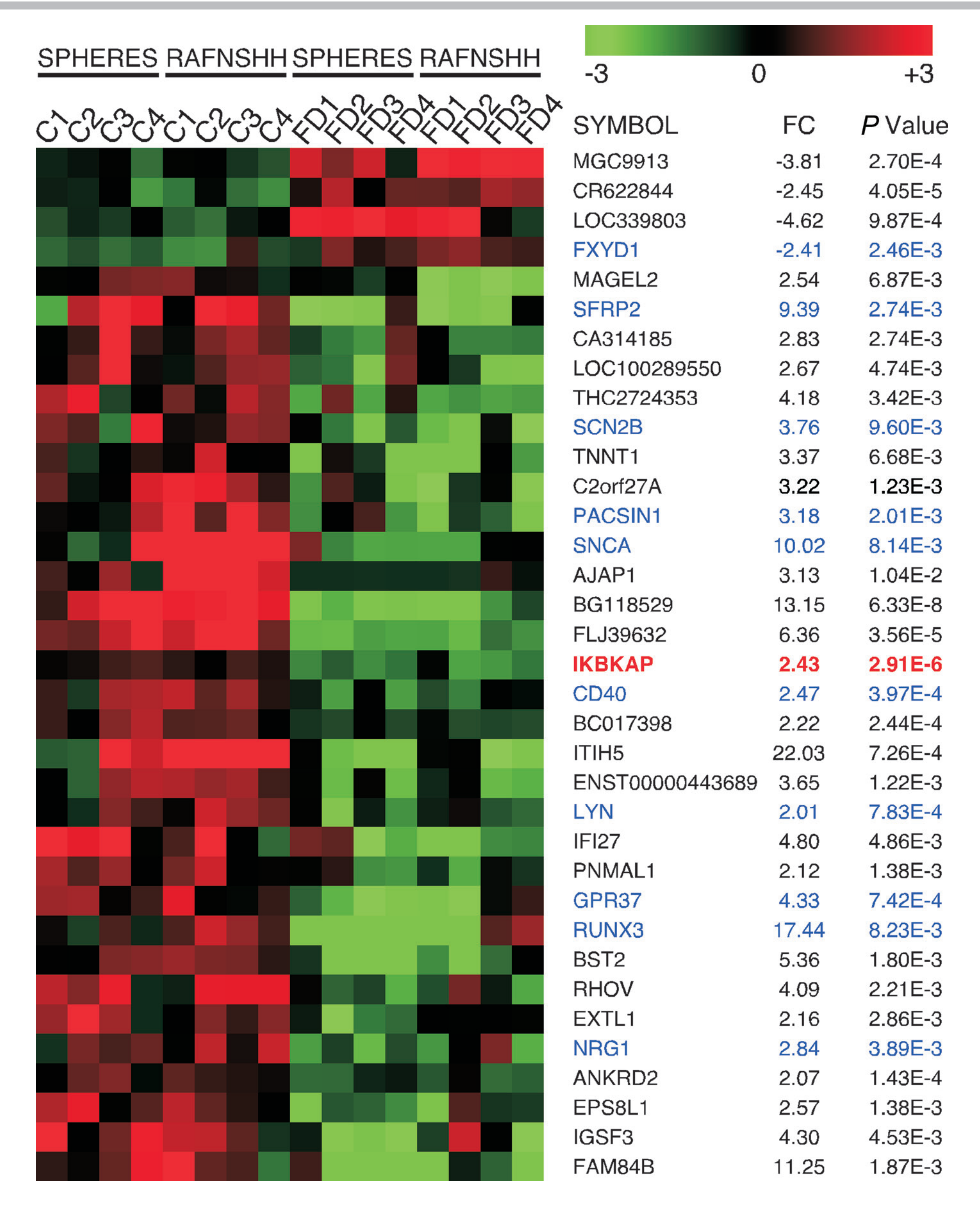


Figure 2. Heatmap of gene expression changes in control versus FD hOE-MSCs. Heatmap representation of overexpressed (red) and underexpressed (green) genes in four controls and four FD OE-MSCs in different culture conditions named as "SPHERES" and "RAFHSHH". Normalized signal intensities were treated with the SAM software to highlight the most differentially expressed genes, with a FDR set at 10%. The color scale bar indicates Log_2 ratio of intensities. Genes related to nervous system development are indicated in blue.

et al., 2006; Marazziti et al., 2007; Newbern and Birchmeier, 2010; Perez-Otano et al., 2006; Tan et al., 2002]. When analyzing the gene ontology (GO) of the dysregulated genes in FD ($P < 0.01$ and FC>2, Supp. Table S2), the pathways with the most significant differential expression correspond to regulation of nervous system development and synaptic vesicle transport (Table 1).

NOVA1 is Differentially Expressed in FD Versus Control Sphere-Derived hOE-MSCs

As previously shown, cells that have been induced to form spheres express a higher amount of *IKBKAP* WT transcript. Therefore, we were interested to identify genes that may be associated with this alternative splicing profile. We were surprised to find that spheres upregulated a significant number of genes related to nervous system development and synaptic transmission (Table 1). Detailed

Table 1. Top Biological Process Gene Ontology (GO) Terms Overrepresented by Dysregulated Genes

Control versus FD cells:

ID	Term	Count	*P* Value	FDR
FUNCTIONAL GROUP 1 ENRICHMENT SCORE: 2.35				
ID:0051960	Regulation of nervous system development	5	1.0E-3	1.6
ID:0050767	Regulation of neurogenesis	4	6.9E-3	10
ID:0060284	Regulation of cell development	4	1.2E-2	17
FUNCTIONAL GROUP 2 ENRICHMENT SCORE: 1.46				
ID:0048489	Synaptic vesicle transport	3	2.7E-3	4.1

Control versus FD sphere cells:

ID	Term	Count	*P* Value	FDR
FUNCTIONAL GROUP 1 ENRICHMENT SCORE: 1.32				
ID:0051960	Regulation of nervous system development	5	1.3E-2	18
FUNCTIONAL GROUP 2 ENRICHMENT SCORE: 1.31				
ID:0007268	Synaptic transmission	6	1.3E-2	19

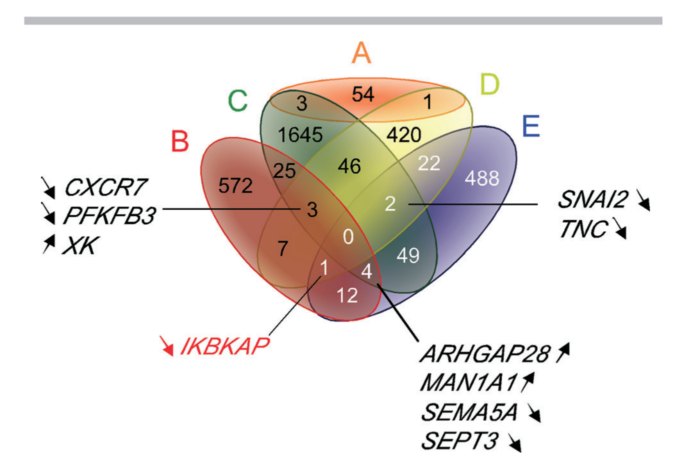


Figure 3. Common genes differentially expressed in FD. Intersection between the current study and the lists of four previous studies for the genes differentially expressed between control and FD/*IKBKAP* knockdown samples (FC>1.5, $P < 0.05$). The genes dysregulated in three different studies are listed and preceded by either a "↘" for underexpression or a "↗" for overexpression in FD samples. Capital letters define each study considered with the following order: **A:** Cheishvili et al. 2007; **B:** Current study; **C:** Lee et al. 2009; **D:** Close et al. 2006; **E:** Cohen-Kupiec et al. 2011.

information about gene expression in spheres is supported in Supp. Table S3. Among nervous system-related genes, we identified genes such as *SNCA* that exhibited a 10-fold downregulation in FD. In addition to finding gene expression alterations for nervous system development in spheres, we also identified *NOVA1* (neuro-oncological ventral antigen 1), encoding a neuron-specific RNA-binding protein [Jelen et al., 2007], as an upregulated gene in FD sphere hOE-MSCs. These results suggest that sphere-forming cells provide an FD-relevant signature even at an early undifferentiated state. Moreover, these results suggest that NOVA1 activity may be involved in the improvement of *IKBKAP* exon 20 inclusion in FD spheres.

Comparative Transcriptome Analysis Identify Convergent Pathways Affected in FD

Four previous studies from other laboratories have generated a wealth of data on the transcriptome variations in either FD or *IKBKAP* knockdown samples [Cheishvili et al., 2007; Close et al., 2006; Cohen-Kupiec et al., 2011; Lee et al., 2009]. Therefore, we procured the raw data from all studies and reanalyzed the data in search for the common candidates that may be involved in FD physiopathology. For each study, we identified genes that are differentially expressed between control and FD/*IKBKAP* knockdown samples with a FC>1.5 and a *P*-value <0.05, and cross-compared the lists of candidate genes for each study (Fig. 3). We did not find genes that were consistently dysregulated in all studies. Among the 3,228 candidate genes differentially expressed in at least one of the five studies, including our own, we found 10 genes shared by three different studies with the same kind of dysregulation. Seven genes were underexpressed in FD (*CXCR7*, *PFKFB3*, *IKBKAP*, *SEMA5A*, *SEPT3*, *SNAI2*, and *TNC*), and three genes were overexpressed (*ARHGAP28*, *MAN1A* and *XK*) (Supp. Table S4). We also analyzed the GO of the 175 genes shared by at least two studies (Supp. Table S4). Nine processes emerged as significantly affected in FD: regulation of cell motion, guanyl ribonucleotide binding, contractile fiber part, neuron differentiation, regulation of protein kinase activity, regulation of apoptosis, cadmium ion binding, muscle tissue development, and osteoblast differentiation (Supp. Table S5).

Kinetin Modulated the Expression of Genes Involved in mRNA Splicing

Our microarray data were next examined for evidence of genes targeted by kinetin. Indeed, this plant cytokinin reproducibly induces rapid increase of *IKBKAP* transcripts with exon 20 inclusion through unknown mechanisms. To further understand the mechanism of kinetin in *IKBKAP* mRNA alternative splicing, we compared FD rafnshh-untreated hOE-MSCs versus FD rafnshh-treated hOE-MSCs with 100 μM of kinetin for 48 hr. Supp. Table S6 displays the list of genes affected by kinetin action in FD rafnshh hOE-MSCs. Interestingly, a majority of candidate genes were downregulated in response to kinetin. In addition to confirming an increased expression of *IKBKAP* in FD hOE-MSCs, we observed cellular responses that are consistent with predicted mechanisms of kinetin action. Indeed, our analysis detected differences in expression of genes involved in mRNA splicing: *LUC7L, SNRPA, WDR70* (Supp. Table S6). Of particular interest, *SNRPA* and *LUC7L* are both related to the U1 snRNP splicing complex required for 5′ss selection. *SNRPA*, downregulated by 1.7-fold in response to kinetin in FD rafnshh-treated hOE-MSCs, encodes the U1 snRNP core protein U1A [Nelissen et al., 1991], *LUC7L*, upregulated by 2-fold, encodes a putative RNA-binding protein similar to the yeast Luc7p subunit of the U1 snRNP [Fortes et al., 1999; Tufarelli et al., 2001].

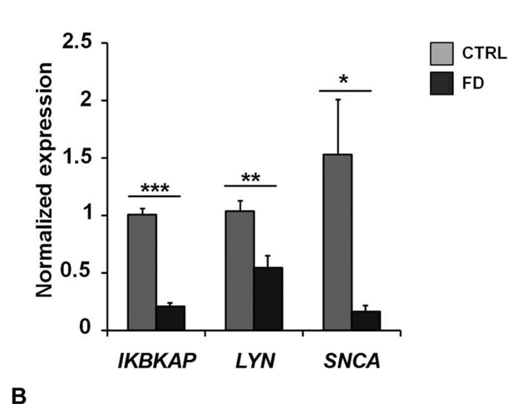
RT-qPCR Analysis of Candidate Genes Validates Microarray Data

To further confirm gene expression data from microarray analysis, we used relative qPCR to verify the differential expression of a subset of the identified genes based on statistical significance, as well as the biological relevance for each comparison. *WDR59* was selected as the reference gene since it exhibited relatively stable expression in our microarray data. Using *IKBKAP* expression as a positive control for each experiment, we confirmed the differential expression of *LYN* and *SNCA* between control and FD cells (Fig. 4A), *MAP1LC3C*, *NOVA1*, *SNCA*, *SPON1* between control and FD sphere-derived cells (Fig. 4B), and *LUC7L* between FD cells with or without kinetin treatment (Fig. 4C).

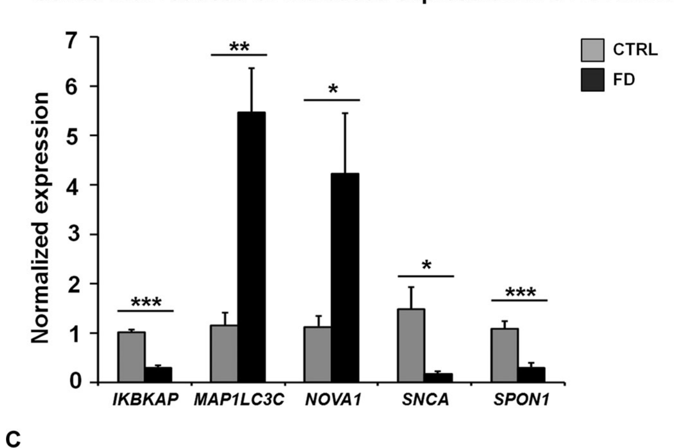
ZNF280D is a Potential Sequence-Specific Target of Kinetin in FD hOE-MSCs

Among the list of genes whose expression is downregulated after kinetin treatment in FD OE-MSCs, we noted the presence of *ZNF280D* (Supp. Table S6). *ZNF280D* belongs to a unique group of 12 genes in the entire genome that contains an alternative 5′ss in one of its exons (exon 16) that is identical to the FD 5′ss (CAAguaagc) [Ibrahim et al., 2007]. Therefore, we hypothesized that kinetin may favor the splicing of introns flanked by the CAAguaagc 5′ss motif, resulting in a modification in the ratio of alternative 5′ss choice for *ZNF280D* exon 16 (Supp. Fig. S2). Since the use of the 5′ss identical to the FD *IKBKAP* intron 20 5′ss is also expected to induce a premature stop codon in *ZNF280D* exon 17 and make it a target for nonsense–mediated mRNA decay (NMD), this may explain why the total amount of *ZNF280D* transcripts is reduced in FD hOE-MSCs after kinetin treatment.

A



B



C

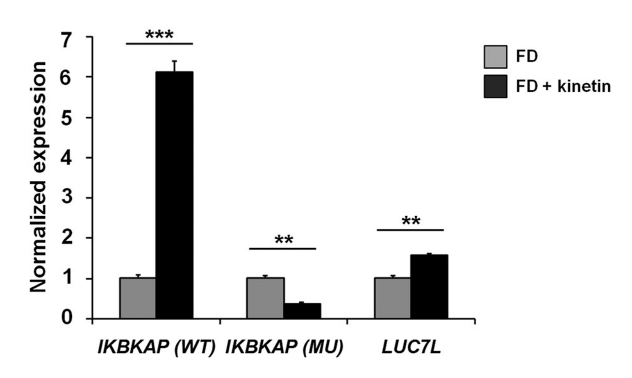


Figure 4. Validation of microarray candidates by RT-qPCR. RT-qPCR using total RNAs extracted from four controls and four FD hOE-MSCs. Histograms represent the mean value of (**A**) *IKBKAP*, *LYN, SNCA*, (**B**) *MAP1LC3C*, *NOVA1*, *SPON1*, and (**C**) *LUC7L* transcript expression level, relative to *WDR59* as a reference gene in control (gray) and FD samples (black). For dysregulated genes between control and FD hOE-MSCs, we pooled values of spheres and differentiated cells for each group. Error bars denote standard errors. (* $P < 0.05$; ** $P < 0.01$, *** $P < 0.001$ using two-tailed Student's test).

IKBKAP, *LUC7L,* and *ZNF280D* are Sensitive to Kinetin Treatment in hOE-MSCs

To corroborate the expression levels of *LUC7L* and *ZNF280D* detected in our microarray hybridization after kinetin treatment, we exposed adherent hOE-MSCs to increasing concentrations of kinetin (25 to 200 μM) over a 48-hr time course and determined the expression changes of these genes relative qPCR (Fig. 5A). Although a dose-dependent action of kinetin on increasing *IKBKAP* WT transcripts was only observed in FD samples (Fig. 5A, higher panel), *LUC7L* expression increased in both control and FD samples (Fig. 5A, middle panel). We also observed that increasing kinetin concentration leads to a dose-dependent inhibition of *ZNF280D* mRNA expression, supporting our hypothesis of sequence-specific targeting by kinetin (Fig. 5A, lower panel). To validate the action of kinetin on the expression of *LUC7L* and *ZNF280D*, we exposed

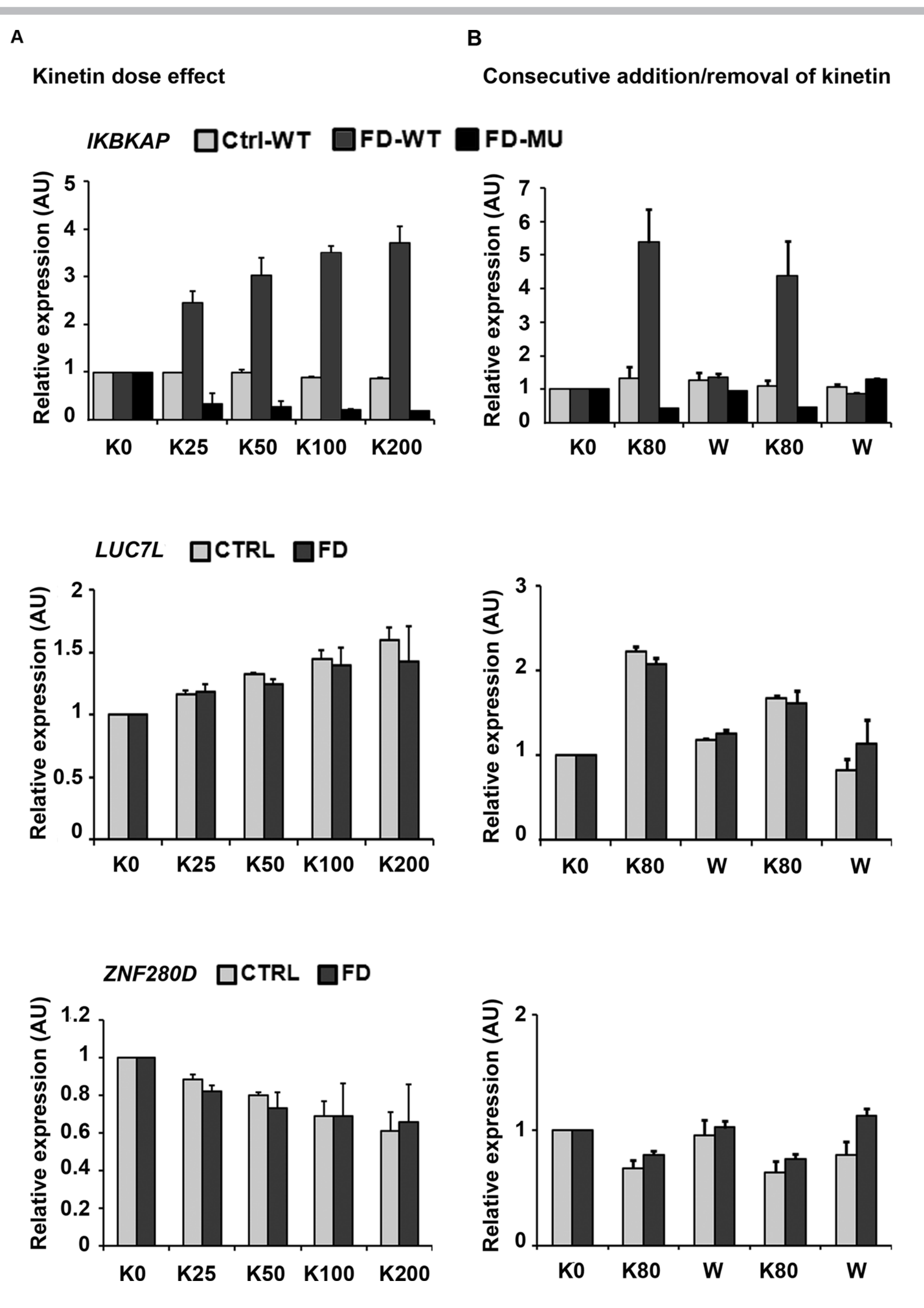


Figure 5. Changes in gene expression after different exposures of hOE-MSCs to kinetin. **A:** Control and FD-adherent hOE-MSCs were incubated for 48 hr with different concentration of kinetin (25, 50, 100, and 200 μM) for dose effect experiment. **B**: Cells were exposed to 80 μM kinetin for 24 hr (K80), followed by the removal of the drug for another 24 hr (i.e., "W" for washout). Two rounds of drug addition/removal were performed and RNA was extracted each time after 24 hr for each condition. Total RNAs were reverse transcribed and levels of expression of *IKBKAP* alternative transcripts as well as *LUC7L* and *ZNF280D* expressions were analyzed by RT-qPCR. Each gene was normalized using *WDR59* as a reference gene.

hOE-MSCs to two consecutive rounds of 24-hr treatment with 80-μM kinetin followed by a 24-hr wash-out. At each 24-hr time point with kinetin treatment, we analyzed gene expression by RT-qPCR and observed that WT *IKBKAP* transcripts and *LUC7L* expression increased while MU *IKBKAP* transcripts and *ZNF280D* expression decreased (Fig. 5B). This variation in expression returned to basal levels during washout period. As expected, in control cells, kinetin treatment modulated expression of *LUC7L* and *ZNF280D* without acting on *IKBKAP* WT isoforms. These results strongly suggest that kinetin may increase the efficiency of 5′ss recognition in the FD context through the recruitment of U1 snRNP.

Genes Involved in mRNA Splicing Display an *IKBKAP*-Like Pattern of Expression

When analyzing gene expression data, it is informative to include a clustering algorithm to find groups of genes that behave similarly over a number of experiments [Eisen et al., 1998; Slonim, 2002]. To

better understand the FD physiopathology and since *IKBKAP* represents the best biomarker to discriminate between control and FD samples, as well as samples with or without kinetin treatment, we wanted to identify genes with expression pattern similar to that of *IKBKAP*. We used hierarchical clustering to create dendrograms that capture the degree of similarity for each gene. An illustrative set of selected genes is shown in Supp. Figure S3A. Next, we looked for the cluster of genes that include *IKBKAP* (Supp. Fig. S3B). Significantly, among the few genes in the same cluster as *IKBKAP*, we identified *DDX42*, which encodes SF3b125, an RNA helicase involved in spliceosome assembly [Will et al., 2002], and *NHP2L1* (nonhistone chromosome protein 2-like 1), which binds the 5'-stem-loop of U4 snRNA and may play a role in late stage spliceosome assembly [Nottrott et al., 1999].

Discussion

Genome-wide expression studies have been widely used in an effort to identify signatures that can define pathologies. In this study, we proposed to use properties of hOE-MSCs to perform a transcriptome analysis of FD. These cells have been used as a nervous system replacement cells in mice [Nivet et al., 2011] and demonstrate a potential to differentiate into nervous cell types [Delorme et al., 2010; Murrell et al., 2005]. Importantly, this novel patient-derived cellular model has allowed us to modulate *IKBKAP* alternative splicing by exposing cells to different culture conditions [Boone et al., 2010]. In this study, we discuss the opportunity to use hOE-MSCs derived from FD patients to analyze the transcriptional differences due to the alteration or improvement of *IKBKAP* mRNA alternative splicing. We focused on identifying gene expression differences in FD using two different cellular models to reproduce neuronal cells in early development (spheres), and neuroglial progenitors in later developmental stages using the "rafnshh" treatment. Retinoic acid (RA) and Sonic hedgehog (Shh) are known to regulate neuronal specification and differentiation during development [Probst et al., 2011]. Both RA and Shh induced expression of a set of genes and proteins that define peripheral nervous system sensory neurons in murine mesenchymal stem cells [Kondo et al., 2005]. These factors were also shown to stimulate the expression of motoneuronal transcription factors in parallel to neurite formation on hOE-MSCs [Zhang et al., 2006].

Previous microarray studies of FD were unable to discriminate *IKBKAP* expression between FD and control cells [Boone et al., 2010; Cheishvili et al., 2007; Keren et al., 2010; Lee et al., 2009]. However, in our analysis, we detected an *IKBKAP* signal above background level in both control and FD patient samples. In addition, we found that *IKBKAP* was the best marker for FD since this gene was initially underexpressed in FD cells but then showed even higher expression after kinetin treatment. These results increased confidence in interpreting our microarray data.

In accordance with previous microarray studies [Boone et al., 2010; Cheishvili et al., 2007; Lee et al., 2009], the FD transcriptional signature is characterized by a general decrease in transcriptional expression that might reflect a defect in transcription elongation due to impaired Elongator activity [Close et al., 2006]. Moreover, gene expression profiling studies have shown that most of gene expression differences between control and FD samples are involved in nervous system development, which correlates with FD physiopathology and findings from other cellular systems [Chen et al., 2009b; Cohen-Kupiec et al., 2011; Lee et al., 2009].

When we explored the transcriptome of spheres, we hypothesized that such cell populations maintained at a higher undifferentiated state would likely reveal discriminating markers of the "stem" state. Interestingly, rather than displaying a profile that is more consistent with stem cells, we identified nervous system-related genes in spheres. In fact, spheres contain a heterogeneous mixture of cells and progenitors whose identity and proportion still need to be characterized. However, this discrepancy with our hypothesis suggests that spheres can be a relevant model for predicting FD alteration, as also proposed for other diseases such as schizophrenia and Parkinson's disease [Cook et al., 2011; Matigian et al., 2010].

As in studies for all rare diseases, the sample size is unavoidably small, which may lead to moderate differences in gene expression variations. In addition, previous investigations, at the genome-wide level, aiming to identify transcriptional defects associated to FD used different cell types. Some investigators treated HeLa or neuroblastoma cells with siRNAs, while others generated FD iPSCs, hOE-MSCs, or analyzed FD brains [Boone et al., 2010; Cheishvili et al., 2007; Close et al., 2006; Cohen-Kupiec et al., 2011; Lee et al., 2009]. It was thus expected from such heterogeneity in cell types, genetic background, and methodologies that important discrepancies would characterize those studies and ours. Despite such limitations, we were able to identify a common set of genes in our microarray data and data from four previous studies (Supp. Table S4), that could contribute to the FD disease process [Cheishvili et al., 2007; Close et al., 2006; Cohen-Kupiec et al., 2011; Lee et al., 2009]. Among the dysregulated genes shared by at least two studies, several are related to nervous system development and characterize common alterations of neuronal cells. Notable downregulated genes include: *SEMA5A* and *SEMA3C*, which encode members of the semaphorin family, involved in axonal guidance during neural development [Hernandez-Montiel et al., 2008; Hilario et al., 2009]; *NRCAM*, which encodes an adhesion molecule acting as a co-receptor for SEMA3B and 3F [Falk et al., 2005]; *ALCAM*, involved in axonal guidance [Buhusi et al., 2009]; *RELN*, which regulates the migration of neuroblasts [Frotscher, 2010]; *FEZ1*, which promotes neurite elongation [Maturana et al., 2010]; and *DLX5*, which encodes a homeobox transcriptional factor promoting neuronal differentiation [Perera et al., 2004]. Therefore, we can speculate that in FD, the dysregulation of these candidate genes in FD will disrupt the precisely defined waves of migration, differentiation, and navigation of axonal growth cone for synapse formation, which are all essential for the formation of the peripheral nervous system. *LYN* is one of the genes that was found to be downregulated in our microarray data, as validated by RT-qPCR and *IKBKAP* knockdown in HeLa cells. *LYN* encodes a Src family tyrosine kinase that have many roles in the process of oligodendrocyte differentiation [Colognato et al., 2004; Hossain et al., 2010], and dopamine release in the mesolimbic system [Gibb et al., 2011]. Importantly, we highlighted 10 genes, including *IKBKAP*, whose dysregulation is shared by three independent genome-wide transcriptional studies (Fig. 3). Notably, four of them, *CXCR7*, *SEMA5A*, *SNAI2*, and *TNC*, are closely related to cell migration [Katafiasz et al., 2011; Nishio et al., 2005; Sadanandam et al., 2010; Sanchez-Alcaniz et al., 2011]. Since several studies previously suggested a contribution of altered migration pathways in the physiopathology of FD [Close et al., 2006; Cohen-Kupiec et al., 2011; Creppe et al., 2009; Johansen et al., 2008; Lee et al., 2009; Naumanen et al., 2008], future experiments will aim to investigate the role of those four genes in functional migration assays using hOE-MSCs.

Understanding the mechanisms underlying regulation of tissue-specific gene expression remains a challenging problem. So far, the only candidate gene that may explain increased aberrant splicing of *IKBKAP* mRNA in the nervous system is *NOVA1*, identified by

Lee et al. as a downregulated gene in FD iPSC-derived neural crest precursors [Lee et al., 2009]. *NOVA1* is a tissue-specific factor regulating alternative splicing in the brain of a large number of genes that function primarily at synapses [Ule et al., 2005]. Thus, it has been suggested that this splicing factor may participate in the balance of neuronal excitation and inhibition, and is necessary for proper synaptic development and function [Ruggiu et al., 2009]. In addition, one of the roles of NOVA proteins may be to enable neurons to adapt their synaptic inhibition in response to neuronal activity [Jelen et al., 2010]. In our system, we confirmed a *NOVA1* dysregulation in FD hOE-MSCs-derived spheres supporting this gene's potentially critical role in modulating *IKBKAP* mRNA alternative splicing. Therefore, we can speculate that *NOVA1* may not only act as a master candidate to regulate *IKBKAP* pre-mRNA splicing in FD, but also the regulation of many other targets involved in progression of this neurodegenerative disease. To understand the precise role of *NOVA1* in mRNA splicing, further experiments modulating its expression in human control and FD cells will be necessary. In addition, it is clear from the initial analysis of *postmortem* tissues that most constitutive *IKBKAP* exon 20 skipping occurs in tissues representing a mixture of cell types, and not just neurons [Cuajungco et al., 2003]. Thus, the ability to derive pure cultures of neurons or glial cells from hOE-MSCs will be of great benefit to determine the cell type predominantly affected during FD development.

We report for the first time a genome-wide gene expression analysis of *IKBKAP* mRNA splicing in response to kinetin, a plant cytokinin. Surprisingly, although kinetin helps to increase WT *IKBKAP* transcript level, the compound does not seem to influence the expression of a large proportion of genes. This specificity in *IKBKAP* mRNA splicing is an encouraging result in light of its potential clinical use [Axelrod et al., 2011]. Although the mechanism by which kinetin improves exon inclusion is still unknown, a previous study has suggested that kinetin may target specific sequences within the 5′ss [Hims et al., 2007]. In this context, our finding that genes encoding a core component and a putative subunit of U1 snRNP, *SNRPA* and *LUC7L*, are regulated by kinetin, supports the hypothesis that this compound can induce the recruitment of splicing factors to reinforce 5′ss recognition. In addition, we demonstrated a consistent decrease of *ZNF280D* expression, which shares with *IKBKAP* an identical 5′ss motif that potentiates the presence of a premature stop codon most likely targeted by the NMD machinery. Therefore, we propose kinetin as a new sequence-specific agent that can affect U1 snRNP-mediated 5′ss recognition. Further experiments considering the 11 other alternatively spliced mRNAs sharing a 5′ss identical to the one bordering *IKBKAP* exon 20 will also be of interest to understand the mechanism underlying kinetin activity on mRNA splicing.

In conclusion, this study provides important clues to the physiopathology of FD. We identified several genes involved in nervous system development and differentiation that could represent the molecular-altered signature unique to the abnormal FD neuronal function. Knowledge of the commonly expressed genes from different cell types should facilitate their further characterization and functional studies. Our results also identified kinetin as a compound that affects genes involved in mRNA maturation and shed new light on its mechanism of action and its potential for therapeutic use.

Acknowledgments

We wish to thank the patients and their families for their contribution to this study. We also thank Jeanne Hsu for critical reading of the manuscript.

References

Anderson SL, Coli R, Daly IW, Kichula EA, Rork MJ, Volpi SA, Ekstein J, Rubin BY. 2001. Familial dysautonomia is caused by mutations of the IKAP gene. Am J Hum Genet 68:753–758.

Aubert J, Dunstan H, Chambers I, Smith A. 2002. Functional gene screening in embryonic stem cells implicates Wnt antagonism in neural differentiation. Nat Biotechnol 20:1240–1245.

Axelrod FB. 2004. Familial dysautonomia. Muscle Nerve 29:352–363.

Axelrod FB, Iyer K, Fish I, Pearson J, Sein ME, Spielholz N. 1981. Progressive sensory loss in familial dysautonomia. Pediatrics 67:517–522.

Axelrod FB, Liebes L, G-vS G, Mendoza S, Mull J, Leyne M, Norcliffe-Kaufmann L, Kaufmann H, Slaugenhaupt SA. 2011. Kinetin improves IKBKAP mRNA splicing in patients with familial dysautonomia. Pediatr Res 70:480–483.

Boone N, Loriod B, Bergon A, Sbai O, Formisano-Treziny C, Gabert J, Khrestchatisky M, Nguyen C, Feron F, Axelrod FB, Ibrahim EC. 2010. Olfactory stem cells, a new cellular model for studying molecular mechanisms underlying familial dysautonomia. PLoS One 5:e15590.

Buhusi M, Demyanenko GP, Jannie KM, Dalal J, Darnell EP, Weiner JA, Maness PF. 2009. ALCAM regulates mediolateral retinotopic mapping in the superior colliculus. J Neurosci 29:15630–15641.

Burré J, Sharma M, Tsetsenis T, Buchman V, Etherton MR, Südhof TC. 2010. Alpha-synuclein promotes SNARE-complex assembly in vivo and in vitro. Science 329:1663–1667.

Cheishvili D, Maayan C, Cohen-Kupiec R, Lefler S, Weil M, Ast G, Razin A. 2011. IKAP/Elp1 involvement in cytoskeleton regulation and implication for familial dysautonomia. Hum Mol Genet 20:1585–1594.

Cheishvili D, Maayan C, Smith Y, Ast G, Razin A. 2007. IKAP/hELP1 deficiency in the cerebrum of familial dysautonomia patients results in down regulation of genes involved in oligodendrocyte differentiation and in myelination. Hum Mol Genet 16:2097–2104.

Chen C, Tuck S, Bystrom AS. 2009a. Defects in tRNA modification associated with neurological and developmental dysfunctions in Caenorhabditis elegans elongator mutants. PLoS Genet 5:e1000561.

Chen YT, Hims MM, Shetty RS, Mull J, Liu L, Leyne M, Slaugenhaupt SA. 2009b. Loss of mouse Ikbkap, a subunit of elongator, leads to transcriptional deficits and embryonic lethality that can be rescued by human IKBKAP. Mol Cell Biol 29:736–744.

Close P, Hawkes N, Cornez I, Creppe C, Lambert CA, Rogister B, Siebenlist U, Merville MP, Slaugenhaupt SA, Bours V, Svejstrup JQ, Chariot A. 2006. Transcription impairment and cell migration defects in elongator-depleted cells: implication for familial dysautonomia. Mol Cell 22:521–531.

Cohen-Kupiec R, Pasmanik-Chor M, Oron-Karni V, Weil M. 2011. Effects of IKAP/hELP1 deficiency on gene expression in differentiating neuroblastoma cells: implications for familial dysautonomia. PLoS One 6:e19147.

Colognato H, Ramachandrappa S, Olsen IM, ffrench-Constant C. 2004. Integrins direct Src family kinases to regulate distinct phases of oligodendrocyte development. J Cell Biol 167:365–375.

Cook AL, Vitale AM, Ravishankar S, Matigian N, Sutherland GT, Shan J, Sutharsan R, Perry C, Silburn PA, Mellick GD, Whitelaw ML, Wells CA, Mackay-Sim A, Wood SA. 2011. NRF2 activation restores disease related metabolic deficiencies in olfactory neurosphere-derived cells from patients with sporadic Parkinson's disease. PLoS One 6:e21907.

Creppe C, Malinouskaya L, Volvert ML, Gillard M, Close P, Malaise O, Laguesse S, Cornez I, Rahmouni S, Ormenese S, Belachew S, Malgrange B, Chapelle JP, Siebenlist U, Moonen G, Chariot A, Nguyen L. 2009. Elongator controls the migration and differentiation of cortical neurons through acetylation of alpha-tubulin. Cell 136:551–564.

Cuajungco MP, Leyne M, Mull J, Gill SP, Lu W, Zagzag D, Axelrod FB, Maayan C, Gusella JF, Slaugenhaupt SA. 2003. Tissue-specific reduction in splicing efficiency of IKBKAP due to the major mutation associated with familial dysautonomia. Am J Hum Genet 72:749–758.

Delorme B, Nivet E, Gaillard J, Haupl T, Ringe J, Deveze A, Magnan J, Sohier J, Khrestchatisky M, Roman FS, Charbord P, Sensebe L, Layrolle P, Feron F. 2010. The human nose harbors a niche of olfactory ectomesenchymal stem cells displaying neurogenic and osteogenic properties. Stem Cells Dev 19:853–866.

Deng V, Matagne V, Banine F, Frerking M, Ohliger P, Budden S, Pevsner J, Dissen GA, Sherman LS, Ojeda SR. 2007. FXYD1 is an MeCP2 target gene overexpressed in the brains of Rett syndrome patients and Mecp2-null mice. Hum Mol Genet 16:640–650.

Dong J, Edelmann L, Bajwa AM, Kornreich R, Desnick RJ. 2002. Familial dysautonomia: detection of the IKBKAP IVS20(+6T–>C) and R696P mutations and frequencies among Ashkenazi Jews. Am J Med Genet 110:253–257.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–8.

Esberg A, Huang B, Johansson MJ, Bystrom AS. 2006. Elevated levels of two tRNA species bypass the requirement for elongator complex in transcription and exocytosis. Mol Cell 24:139–148.

Falk J, Bechara A, Fiore R, Nawabi H, Zhou H, Hoyo-Becerra C, Bozon M, Rougon G, Grumet M, Puschel AW, Sanes JR, Castellani V. 2005. Dual functional activity of semaphorin 3B is required for positioning the anterior commissure. Neuron 48:63–75.

Fortes P, Bilbao-Cortes D, Fornerod M, Rigaut G, Raymond W, Seraphin B, Mattaj IW. 1999. Luc7p, a novel yeast U1 snRNP protein with a role in 5' splice site recognition. Genes Dev 13:2425–2438.

Frotscher M. 2010. Role for Reelin in stabilizing cortical architecture. Trends Neurosci 33:407–414.

Gibb SL, Jeanblanc J, Barak S, Yowell QV, Yaka R, Ron D. 2011. Lyn kinase regulates mesolimbic dopamine release: implication for alcohol reward. J Neurosci 31:2180–2187.

Hawkes NA, Otero G, Winkler GS, Marshall N, Dahmus ME, Krappmann D, Scheidereit C, Thomas CL, Schiavo G, Erdjument-Bromage H, Tempst P, Svejstrup JQ. 2002. Purification and characterization of the human elongator complex. J Biol Chem 277:3047–3052.

Hernandez-Montiel HL, Tamariz E, Sandoval-Minero MT, Varela-Echavarria A. 2008. Semaphorins 3A, 3C, and 3F in mesencephalic dopaminergic axon pathfinding. J Comp Neurol 506:387–397.

Hilario JD, Rodino-Klapac LR, Wang C, Beattie CE. 2009. Semaphorin 5A is a bifunctional axon guidance cue for axial motoneurons in vivo. Dev Biol 326:190–200.

Hims MM, Ibrahim EC, Leyne M, Mull J, Liu L, Lazaro C, Shetty RS, Gill S, Gusella JF, Reed R, Slaugenhaupt SA. 2007. Therapeutic potential and mechanism of kinetin as a treatment for the human splicing disease familial dysautonomia. J Mol Med 85:149–161.

Hossain S, Fragoso G, Mushynski WE, Almazan G. 2010. Regulation of peripheral myelination by Src-like kinases. Exp Neurol 226:47–57.

Huang B, Johansson MJ, Bystrom AS. 2005. An early step in wobble uridine tRNA modification requires the Elongator complex. RNA 11:424–436.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57.

Ibrahim EC, Hims MM, Shomron N, Burge CB, Slaugenhaupt SA, Reed R. 2007. Weak definition of IKBKAP exon 20 leads to aberrant splicing in familial dysautonomia. Hum Mutat 28:41–53.

Jelen N, Ule J, Zivin M. 2010. Cholinergic regulation of striatal Nova mRNAs. Neuroscience 169:619–627.

Jelen N, Ule J, Zivin M, Darnell RB. 2007. Evolution of Nova-dependent splicing regulation in the brain. PLoS Genet 3:1838–1847.

Johansen LD, Naumanen T, Knudsen A, Westerlund N, Gromova I, Junttila M, Nielsen C, Bottzauw T, Tolkovsky A, Westermarck J, Coffey ET, Jaattela M, Kallunki T. 2008. IKAP localizes to membrane ruffles with filamin A and regulates actin cytoskeleton organization and cell migration. J Cell Sci 121:854–864.

Katafiasz D, Smith LM, Wahl JK, 3rd. 2011. Slug (SNAI2) expression in oral SCC cells results in altered cell-cell adhesion and increased motility. Cell Adhes Migr 5:315–322.

Keren H, Donyo M, Zeevi D, Maayan C, Pupko T, Ast G. 2010. Phosphatidylserine increases IKBKAP levels in familial dysautonomia cells. PLoS One 5:e15884.

Kondo T, Johnson SA, Yoder MC, Romand R, Hashino E. 2005. Sonic hedghog and retinoic acid synergistically promote sensory fate specification from bone marrow-derived pluripotent stem cells. Proc Natl Acad Sci USA 102:4789–4794.

Kramer I, Sigrist M, de Nooij JC, Taniuchi I, Jessell TM, Arber S. 2006. A role for Runx transcription facto signaling in dorsal root ganglion sensory neuron diversification. Neuron 49:379–393.

Lee G, Papapetrou EP, Kim H, Chambers SM, Tomishima MJ, Fasano CA, Ganat YM, Menon J, Shimizu F, Viale A, Tabar V, Sadelain M, Studer L. 2009. Modelling pathogenesis and treatment of familial dysautonomia using patient-specific iPSCs. Nature 461:402–406.

Li Q, Fazly AM, Zhou H, Huang S, Zhang Z, Stillman B. 2009. The elongator complex interacts with PCNA and modulates transcriptional silencing and sensitivity to DNA damage agents. PLoS Genet 5:e1000684.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25;402–408.

Lopez-Santiago LF, Pertin M, Morisod X, Chen C, Hong S, Wiley J, Decosterd I, Isom LL. 2006. Sodium channel beta2 subunits regulate tetrodotoxin-sensitive sodium channels in small dorsal root ganglion neurons and modulate the response to pain. J Neurosci 26:7984–7994.

Marazziti D, Mandillo S, Di Pietro C, Golini E, Matteoni R, Tocchini-Valentini GP. 2007. GPR37 associates with the dopamine transporter to modulate dopamine uptake and behavioral responses to dopaminergic drugs. Proc Natl Acad Sci USA 104:9846–9851.

Matigian N, Abrahamsen G, Sutharsan R, Cook AL, Vitale AM, Nouwens A, Bellette B, An J, Anderson M, Beckhouse AG, Bennebroek M, Cecil R, Chalk AM, Cochrane J, Fan Y, Feron F, McCurdy R, McGrath JJ, Murrell W, Perry C, Raju J, Ravishankar S, Silburn PA, Sutherland GT, Mahler S, Mellick GD, Wood SA, Sue CM, Wells CA, Mackay-Sim A. 2010. Disease-specific, neurosphere-derived cells as models for brain disorders. Dis Model Mech 3:785–798.

Maturana AD, Fujita T, Kuroda S. 2010. Functions of fasciculation and elongation protein zeta-1 (FEZ1) in the brain. Sci World J 10:1646–1654.

Murrell W, Feron F, Wetzig A, Cameron N, Splatt K, Bellette B, Bianco J, Perry C, Lee G, Mackay-Sim A. 2005. Multipotent stem cells from adult olfactory mucosa. Dev Dynam 233:496–515.

Naumanen T, Johansen LD, Coffey ET, Kallunki T. 2008. Loss-of-function of IKAP/ELP1: could neuronal migration defect underlie familial dysautonomia? Cell Adhes Migr 2:236–239.

Nelissen RL, Sillekens PT, Beijer RP, Geurts van Kessel AH, van Venrooij WJ. 1991. Structure, chromosomal localization and evolutionary conservation of the gene encoding human U1 snRNP-specific A protein. Gene 102:189–196.

Newbern J, Birchmeier C. 2010. Nrg1/ErbB signaling networks in Schwann cell development and myelination. Semin Cell Dev Biol 21:922–928.

Nishio T, Kawaguchi S, Yamamoto M, Iseda T, Kawasaki T, Hase T. 2005. Tenascin-C regulates proliferation and migration of cultured astrocytes in a scratch wound assay. Neuroscience 132:87–102.

Nivet E, Vignes M, Girard SD, Pierrisnard C, Baril N, Deveze A, Magnan J, Lante F, Khrestchatisky M, Feron F, Roman FS. 2011. Engraftment of human nasal olfactory stem cells restores neuroplasticity in mice with hippocampal lesions. J Clin Invest 121:2808–2820.

Nottrott S, Hartmuth K, Fabrizio P, Urlaub H, Vidovic I, Ficner R, Luhrmann R. 1999. Functional interaction of a novel 15.5kD [U4/U6.U5] tri-snRNP protein with the 5' stem-loop of U4 snRNA. EMBO J 18:6119–6133.

Okada Y, Yamagata K, Hong K, Wakayama T, Zhang Y. 2010. A role for the elongator complex in zygotic paternal genome demethylation. Nature 463:554–558.

Pearson J, Pytel BA. 1978. Quantitative studies of sympathetic ganglia and spinal cord intermedio-lateral gray columns in familial dysautonomia. J Neurol Sci 39:47–59.

Pearson J, Pytel BA, Grover-Johnson N, Axelrod F, Dancis J. 1978. Quantitative studies of dorsal root ganglia and neuropathologic observations on spinal cords in familial dysautonomia. J Neurol Sci 35:77–92.

Perera M, Merlo GR, Verardo S, Paleari L, Corte G, Levi G. 2004. Defective neuronogenesis in the absence of Dlx5. Mol Cell Neurosci 25:153–161.

Perez-Otano I, Lujan R, Tavalin SJ, Plomann M, Modregger J, Liu XB, Jones EG, Heinemann SF, Lo DC, Ehlers MD. 2006. Endocytosis and synaptic removal of NR3A-containing NMDA receptors by PACSIN1/syndapin1. Nat Neurosci 9:611–621.

Probst S, Kraemer C, Demougin P, Sheth R, Martin GR, Shiratori H, Hamada H, Iber D, Zeller R, Zuniga A. 2011. SHH propagates distal limb bud development by enhancing CYP26B1-mediated retinoic acid clearance via AER-FGF signalling. Development 138:1913–1923.

Rahl PB, Chen CZ, Collins RN. 2005. Elp1p, the yeast homolog of the FD disease syndrome protein, negatively regulates exocytosis independently of transcriptional elongation. Mol Cell 17:841–853.

Ruggiu M, Herbst R, Kim N, Jevsek M, Fak JJ, Mann MA, Fischbach G, Burden SJ, Darnell RB. 2009. Rescuing Z+ agrin splicing in Nova null mice restores synapse formation and unmasks a physiologic defect in motor neuron firing. Proc Natl Acad Sci USA 106:3513–3518.

Sadanandam A, Rosenbaugh EG, Singh S, Varney M, Singh RK. 2010. Semaphorin 5A promotes angiogenesis by increasing endothelial cell proliferation, migration, and decreasing apoptosis. Microvasc Res 79:1–9.

Sanchez-Alcaniz JA, Haege S, Mueller W, Pla R, Mackay F, Schulz S, Lopez-Bendito G, Stumm R, Marin O. 2011. Cxcr7 controls neuronal migration by regulating chemokine responsiveness. Neuron 69:77–90.

Scott SA, Edelmann L, Liu L, Luo M, Desnick RJ, Kornreich R. 2010. Experience with carrier screening and prenatal diagnosis for 16 Ashkenazi Jewish genetic diseases. Hum Mutat 31:1240–1250.

Shetty RS, Gallagher CS, Chen YT, Hims MM, Mull J, Leyne M, Pickel J, Kwok D, Slaugenhaupt SA. 2011. Specific correction of a splice defect in brain by nutritional supplementation. Hum Mol Genet 20:4093–4101.

Slaugenhaupt SA, Blumenfeld A, Gill SP, Leyne M, Mull J, Cuajungco MP, Liebert CB, Chadwick B, Idelson M, Reznik L, Robbins C, Makalowska I, Brownstein M, Krappmann D, Scheidereit C, Maayan C, Axelrod FB, Gusella JF. 2001. Tissue-specific expression of a splicing mutation in the IKBKAP gene causes familial dysautonomia. Am J Hum Genet 68:598–605.

Slaugenhaupt SA, Mull J, Leyne M, Cuajungco MP, Gill SP, Hims MM, Quintero F, Axelrod FB, Gusella JF. 2004. Rescue of a human mRNA splicing defect by the plant cytokinin kinetin. Hum Mol Genet 13:429–436.

Slonim DK. 2002. From patterns to pathways: gene expression data analysis comes of age. Nat Genet (Suppl 32):502–508.

Solinger JA, Paolinelli R, Kloss H, Scorza FB, Marchesi S, Sauder U, Mitsushima D, Capuani F, Sturzenbaum SR, Cassata G. 2010. The Caenorhabditis elegans Elongator complex regulates neuronal alpha-tubulin acetylation. PLoS Genet 6:e1000820.

Tan J, Town T, Mori T, Obregon D, Wu Y, DelleDonne A, Rojiani A, Crawford F, Flavell RA, Mullan M. 2002. CD40 is expressed and functional on neuronal cells. EMBO J 21:643–652.
Tufarelli C, Frischauf AM, Hardison R, Flint J, Higgs DR. 2001. Characterization of a widely expressed gene (LUC7-LIKE; LUC7L) defining the centromeric boundary of the human alpha-globin domain. Genomics 71:307–314.
Ule J, Ule A, Spencer J, Williams A, Hu JS, Cline M, Wang H, Clark T, Fraser C, Ruggiu M, Zeeberg BR, Kane D, Weinstein JN, Blume J, Darnell RB. 2005. Nova regulates brain-specific splicing to shape the synapse. Nat Genet 37:844–852.
Will CL, Urlaub H, Achsel T, Gentzel M, Wilm M, Luhrmann R. 2002. Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5p homologue and an SF3b DEAD-box protein. EMBO J 21:4978–4988.
Zhang X, Klueber KM, Guo Z, Cai J, Lu C, Winstead WI, Qiu M, Roisen FJ. 2006. Induction of neuronal differentiation of adult human olfactory neuroepithelial-derived progenitors. Brain Res 1073–1074:109–119.

3.5 Conclusions et perspectives

Les puces à ADN, une technique très importante pour l'étude des pathologies

L'analyse de données de puces à ADN reste une technique de choix pour l'étude de l'expression des ARN messagers. En effet, cette technique a révolutionné l'étude des pathologies et a permis de grandes avancées au niveau de la compréhension des mécanismes, du développement et du traitement de ces maladies, bien que cela ne soit pas toujours évident, et nécessite parfois l'utilisation de tests statistiques ou d'approches parfois plus contestées, lors d'études où il n'y a pas voir peu de réplicats par exemple.

Comme les analyses effectuées dans ce chapitre le montre, elle permet l'obtention de signatures moléculaires de bonne qualité et reproductible que ce soit dans le cadre de maladies mono-génique, comme la Dysautonomie Familiale, ou dans le cas de maladies plus complexes comme les maladies infectieuses telles que la dengue. Cependant, il est toutefois à noter que selon la pathologie étudié, ces signatures sont plus ou moins étendues, et nécessite parfois des expériences supplémentaires.

Des techniques à haut débit à celles à très haut débit

Le développement récent des techniques de séquençage à très haut débit et les nombreuses découvertes concernant la régulation de l'expression des gènes comme le rôle des miRNA, des lincRNA ont permis l'évolution des puces à ADN pour proposer maintenant l'étude de ces ARN non codants. Des puces dédiées à l'étude des milliers de miRNA découvert dans le génome humain ont été créées, alors que les lincRNA présents en plus petit nombre (environ 200) ont été tout simplement rajoutés aux puces à ADN d'expression.

On s'oriente donc vers des conceptions d'études plus complexes avec en plus des puces à ADN, des puces à miRNA. En effet, les ARN non-codant ont été identifiés comme des cibles thérapeutiques potentielles pouvant traiter des maladies complexes comme le cancer. Leur étude est donc un enjeu thérapeutique mais également commerciale.

Cependant, il existe quelques limitations des puces à ADN par rapport au séquençage à très haut débit. En effet, les techniques de séquençage à très haut débit sont plus sensibles et permettent d'avoir accès à plus d'informations concernant les gènes transcrits. En effet, ces dernières permettent de faire du « *tag counting* », de rechercher des transcrits alternatifs ou bien encore de détecter des fusions de gène ([Maher *et al.*, 2009]).

Mais vu son coût, les puces à ADN sont encore de nos jours privilégiées lors d'étude ayant pour but l'identification de signature transcriptionnelle spécifique d'une pathologie ou l'étude de l'effet d'une thérapie. En effet, les puces à ADN permettent d'analyser plus d'échantillons, ce qui est très utile pour des études épidémiologique où l'on dispose de beaucoup d'échantillons. Et enfin, ces données sont beaucoup plus simple à analyser et sont générées très rapidement.

Vers la construction de réseaux de régulation

Une fois l'analyse de données générée, on dispose souvent de liste de gènes sur- et sous-exprimées dans une des conditions. Une fois ces données contextualisées, l'étape suivante de l'analyse consiste en la génération de réseaux de gènes. Ces graphes nécessitent l'intégration, en plus des données de régulation transcriptionnelle disponible, d'autres types de données comme les interactions protéine-protéine par exemple.

Chapitre 4

Fouille de données de puces à ADN

Sommaire

Avec l'augmentation spectaculaire de l'étude du transcriptome par puces à ADN, il est devenu indispensable de stocker les informations relatives aux expériences afin que celle-ci puissent être ré-analysées ou combinées dans le contexte de méta-analyses.

Aujourd'hui, la problématique de ré-analyse et/ou de méta-analyse s'inscrit dans un contexte de génomique intégrative, qui vise ultimement à modéliser le vivant. Cette approche ambitieuse a pour but d'identifier les gènes impliqués dans un processus biologique particulier, de comprendre comment ceux-ci interagissent et de mettre en évidence des perturbations pouvant conduire à un état pathologique. Ceci nécessite a minima une caractérisation du génome, du protéome et du métabolome de l'organisme. L'inférence de réseaux fait appel à des données pouvant être très hétérogènes : données d'expression, d'interactions génétiques ou physiques. Cette modélisation peut être réalisée, notamment, à l'aide du logiciel GINsim développé au sein de notre laboratoire ([Chaouiya *et al.*, 2012]). Afin de pouvoir générer des données de co-expression de gènes à partir de données publiques, le logiciel de méta-analse TranscriptomeBrowser a été créé ([Lopez *et al.*, 2008]). J'ai été fortement impliquée dans ce projet dès mon master.

4.1 Stockage des données

Une fois générées, les données de puces à ADN doivent être stockées puis mis à disposition de la communauté scientifique pour permettre leur ré-analyse. Pour cela, les laboratoires mettent généralement en place des systèmes internes afin de tracer les données, concourant ainsi à leur qualité (LIMS, base de données). Le moyen le plus efficace pour stocker et extraire des informations est l'utilisation de bases de données, système qui a été choisi pour le stockage à long terme des informations et données des expériences de puces à ADN. La principale technologie de base de données couramment utilisée à l'heure actuel est MySQL.

4.1.1 Qualité et traçage

La qualité des données dépend beaucoup de la capacité des utilisateurs à visualiser les problèmes et à apprendre de leurs erreurs par des améliorations constantes des outils et techniques.

Dans cet effort de poursuite des développements bioinformatiques, la traçabilité des données passe par le maintien des suivis de modifications des programmes et/ou des scripts grâce à des logiciels de gestion de versions, comme SVN (subversion) par exemple. Ces logiciels de suivis permettent de mutualiser un développement, en stockant le code source d'un logiciel et une arborescence de fichiers en conservant la chronologie de toutes les modifications qui ont été effectuées. Le système travaille par fusion de copies locales et distantes, et non par écrasement de la version distante par la version locale.

La mise en place d'un système de gestion de l'information du laboratoire (LIMS, « *Laboratory Information Management System* »), logiciel de gestion intégré, permet des actions telles que : la traçabilité des échantillons, la gestion des utilisateurs, des instruments, des stocks, des approvisionnements par fournisseur, le suivi des produits et équipements utilisés, l'enregistrement des incidents, la définition des analyses effectuées avec leurs paramètres, et

parfois même le contrôle à distance des instruments.

4.1.2 Les bases de données MySQL

MySQL est un système de gestion de bases de données (SGBD) permettant d'optimiser les recherches, les tries et la visualisation de quantités importantes de données. Il fait partie des logiciels de gestion de bases de données les plus utilisés dans le monde avec Oracle et Microsoft SQL Server. MySQL est un serveur de bases de données relationnelles utilisant le langage de requête SQL (*Structured Query Language*) développé dans un souci de performances élevées en lecture, ce qui signifie qu'il est davantage orienté vers le service de données déjà en place que vers celui de mises à jour fréquentes et fortement sécurisées. Il est multi-tâches (multi-thread) et multi-utilisateurs. Il fonctionne sous les principaux systèmes d'exploitation et les données qu'il contient sont accessibles en utilisant de très nombreux langages de programmation parmi lesquels Java, C, Perl, PHP ; une interface de programmation (Application Programming Interface ou API) spécifique est disponible pour chacun d'entre eux.

L'une des spécificités de MySQL est de pouvoir gérer plusieurs moteurs de stockage au sein d'une seule base. Chaque table peut utiliser un moteur différent au sein d'une base. La facilité d'utilisation de plusieurs moteurs de stockage dans une seule base permet une énorme flexibilité dans l'optimisation de la base pour chaque table : on utilisera MyISAM pour stocker les données peu sensibles et nécessitant un accès rapide, InnoBD pour les fonctions avancées et les données plus sensibles, MEMORY pour des données pouvant être perdues lors du redémarrage de la machine et souvent modifiées, ARCHIVE pour un historique ne nécessitant que peu de lecture. Les deux principaux moteurs utilisés sont MyISAM et InnoDB. Leur choix est dicté par le type d'application que l'utilisateur veut développer.

4.1.3 Optimisations de base de données

L'optimisation d'une base de données peut se faire à deux niveaux : celui du serveur et celui de la base elle-même. Au niveau du serveur, il est possible de modifier les variables globales et celles de session. Ceci permet d'ajuster l'accessibilité de la base en fonction des applications.

La base de données peut être optimisée par la normalisation/dénormalisation des tables, l'utilisation de jointures et d'index, le découpage de requêtes en requêtes simples et l'utilisation de programmes stockés.

L'utilisation de programmes stockés permet (1) d'augmenter la sécurité de la base de données, (2) d'extraire des données en routine, et (3) de diminuer le trafic sur le réseau.

Il en existe 3 types :
– les **procédures stockées** qui permettent une action mais qui ne retournent aucun résultat ;
– les **fonctions stockées** qui retournent un résultat et peuvent être directement utilisées dans des requêtes comme par exemple la fonction de MySQL DATE() ;

- les **triggers** qui surveillent la base de données et déclenchent des requêtes automatiques, telles que le remplissage de tables à partir d'une insertion de données.

4.2 Méta-analyse et intégration de données

4.2.1 Bases de données de biologie

Pour permettre le stockage et l'organisation des données biologiques à différents niveaux, de nombreuses bases de données ont été mise en place telles que des bases de données de :

- **séquences** : GenBank, EMBL Nucleotide Sequence Database et DNA Data Bank of Japan (DDBJ), Eucaryotic Promoter Database (EPD) ;
- **protéines** : UniProt, Protein Data Bank (PDB), InterPro, Institut européen de bio-informatique (EBI) ;
- **génomiques spécialisées** : Saccharomyces Genome Database (SGD), FlyBase, Worm-Base, The Arabidopsis Information Resource, Zebrafish Information Network ;
- **facteurs de transcription** : TRANFAC, JASPAR ;
- **polymorphismes génétique** : dbSNP, Hapmap ;
- **voies de signalisation** : KEGG, REACTOME, Panther, NCBICP.

Toutes ces bases sont interconnectées grâce à l'utilisation d'un identifiant unique pour caractériser une séquence, un gène, un transcrit ou une protéine, comme c'est le cas sur le site du National Center for Biotechnology Information (NCBI).

Ces bases de données peuvent être accessibles via des sites web (protocole http ou ftp), des API, des librairies R. De plus, certains de ces données sont mises à disposition sous la forme de fichiers plats, c'est-à-dire des fichiers texte tabulés.

4.2.2 Bases de données dédiées aux données de puces à ADN

Pour les puces à ADN, des bases de données ont également été développées. Pour stocker ces données, des standards ont été définis ([Stoeckert *et al.*, 2002]). Certains mettent même à disposition leurs données sous la forme de fichiers plats. Ainsi, le consortium MGED devenu FGED pour Functional Genomics Data Society a développé le standard MIAME pour Minimum Information About a Microarray Experiment en juillet 2010 ([Brazma *et al.*, 2001]). MIAME décrit tous les éléments nécessaires à l'interprétation des résultats et les paramètres expérimentaux utiles pour reproduire une expérience, tels que : (1) les fichiers de données brutes obtenus à partir du scanner, (2) la matrice d'expression normalisée, (3) l'annotation complète des échantillons (type, dose de traitement, temps de cinétique), (4) la conception de l'expérience incluant les relations entre échantillons (réplicats biologiques ou techniques, échantillons ayant subit une différenciation, etc), (5) le type de plateforme de puce à ADN utilisée, et (6) le laboratoire ayant réalisé les expériences et les protocoles utilisés. Cependant le standard MIAME n'impose aucun format de fichier, mais préconise l'utilisation du format MAGE-ML (*MicroArray Gene Expression Markup Language*) qui est un format tabulé.

Il existe beaucoup de bases de données de puces à ADN plus ou moins spécialisées. Un très bon récapitulatif de ces bases de données a été créé par Sophie Lemoine de la plateforme transcriptome de l'Ecole Normale Supérieure (ENS) et est accessible à l'adresse : http://transcriptome.ens.fr/sgdb/tools/data_management.php. Les principales bases de données utilisées sont : *Gene Expression Omnibus* (GEO) ([Edgar *et al.*, 2002], [Barrett *et al.*, 2005], [Wilhite & Barrett, 2012]) au NCBI (États-Unis) et *ArrayExpres* à l'EBI (Angleterre) ([Brazma *et al.*, 2003], [Parkinson *et al.*, 2011]). La quantité de données dans ces bases croit très fortement car il est maintenant obligatoire pour une publication des résultats obtenus de déposer les données brutes et normalisées dans ces bases de données, ceci afin de permettre leur ré-analyse éventuelle.
Des développements sous R ont également été réalisés pour extraire les données contenues dans GEO et ArrayExpress et permettre leur ré-analyse ; ces librairies R sont GEOquery ([Sean & Meltzer, 2007]) et ArrayExpress ([Kauffmann *et al.*, 2009]) respectivement.

4.2.3 Structure des données dans Gene Expression Omnibus (GEO)

Dans GEO (http://www.ncbi.nlm.nih.gov/geo), les données sont regroupées en plateformes de puces à ADN, échantillons et expériences. Une plateforme est composée d'un résumé décrivant la conception de la puce et d'un tableau d'annotation complet des séquences qui y sont fixées. Chaque plateforme est associée à un identifiant unique (GPLxxx). Une plateforme peut être associée à de nombreux échantillons provenant de diverses expériences menées dans des laboratoires indépendants. Pour chaque échantillon associé à un identifiant unique (GSMxxx), les conditions dans lesquelles celui-ci a été obtenu sont décrites suivant les informations requises par le standard MIAME. Un échantillon est référencé dans une seule plateforme mais peut être inclus dans plusieurs expériences. Une expérience (GSExxx) est constituée d'un ensemble d'échantillons, et décrit précisément les paramètres expérimentaux des différents échantillons pour connaître le but de l'étude.

4.2.4 Ré-analyses et méta-analyses de jeux de données provenant de GEO

Diverses approches et outils ont été développés pour permettre la ré-analyse et/ou la méta-analyse des jeux de données disponibles dans les bases de données de puces à ADN. En effet, GEO propose également une visualisation originale des données sous la forme de jeux de données (datasets, GDSxxx) qui représentent des échantillons statistiquement et biologiquement comparables et manuellement vérifiés par les opérateurs de GEO. Ainsi, GEO propose deux type d'outils : GEO Profiles (Figure 4.1 A,B) et GEO Datasets (Figure 4.1 C,D) ([Barrett *et al.*, 2005]). Le premier permet, gène à gène, de visualiser son profil d'expression à travers les échantillons liés à l'expérience (Figure 4.1 B). Le second met à la disposition des utilisateurs des classifications pré-calculées de gènes et d'échantillons pour un nombre important d'expériences tout en proposant des outils de sélection par analyse supervisées (Figure 4.1 D). Cependant, ces outils restent limités, tant du point de vue de la recherche de l'information, que de sa représentation et de son interprétation.

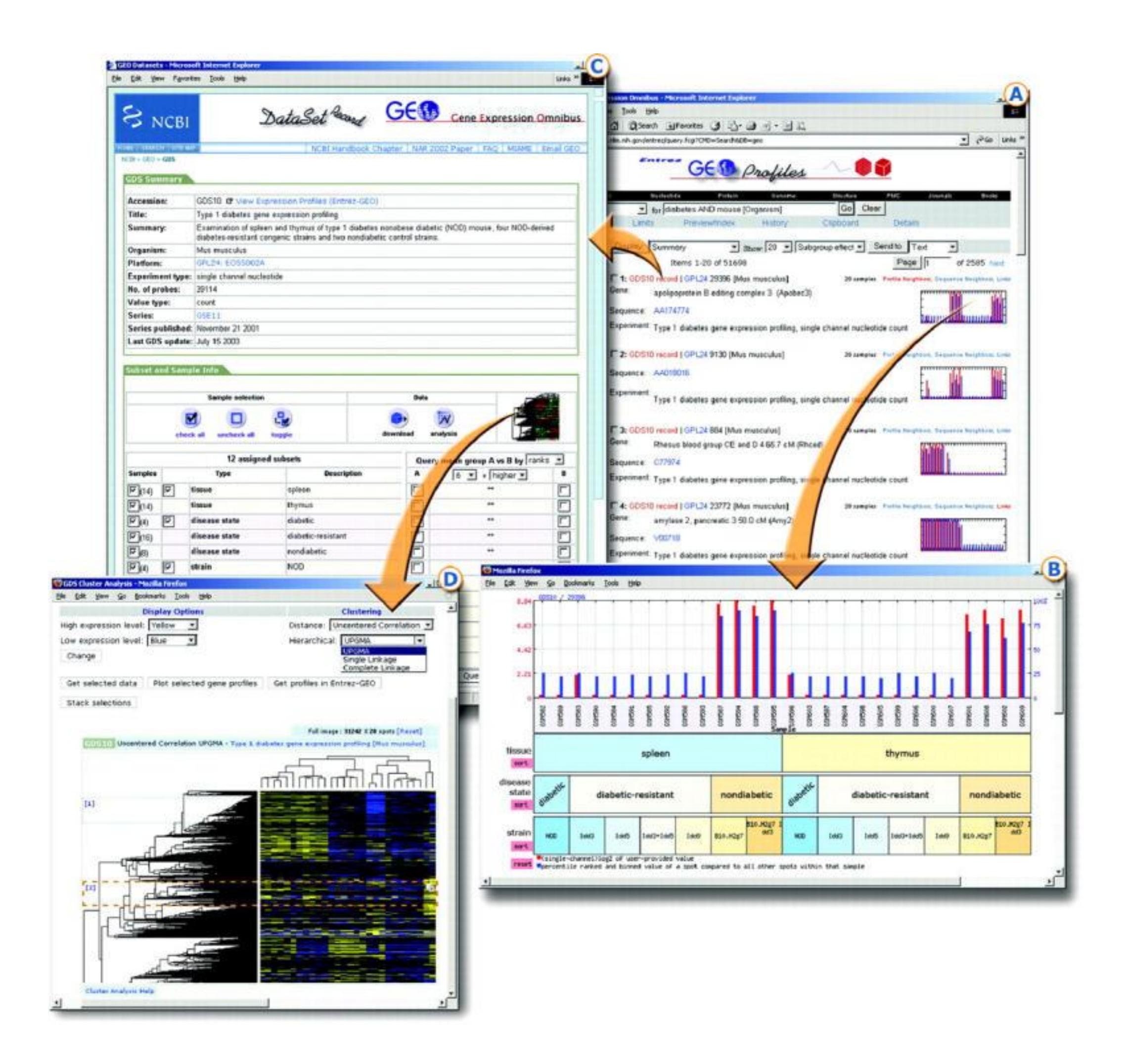


FIGURE 4.1 – Interface web de Gene Expression Omnibus (GEO). (A) GEO Profiles permet de retrouver des expériences ou des jeux de données, (B) représentation sous forme d'histogramme des profils d'expression d'un gène avec en rouge la valeur d'intensité d'expression normalisée et en bleu le rang du gène au sein de chaque échantillon de l'expérience ; chaque échantillon est également mis en relation avec les paramètres expérimentaux (tissu, état de la maladie), (C) résumé des informations sur l'expérience (D) classification hiérarchique du jeu de données (Adapté de [Barrett *et al.*, 2005])

4.3 Contexte du projet

Comme nous l'avons vu, l'utilisation intensive des puces à ADN pour l'étude du transcriptome génère une quantité importante de données. Celles-ci sont depuis quelques années accessibles publiquement par le biais de bases de données en ligne, telles que « *Gene Expression Omnibus* » (GEO, NCBI). La taille de ces bases de données croît très rapidement, et nécessite la mise au point de stratégies d'analyse permettant de ré-analyser ces données efficacement.

Dans ce contexte, j'ai contribué au développement de notre propre outil appelé TranscriptomeBrowser (TBrowser) sous la direction de Denis Puthier en 2007. Le site web du projet est accessible à l'adresse : http://tagc.univ-mrs.fr/tbrowser/
Il n'existait pas d'outils permettant la méta-analyse de jeux de données. Seuls les outils disponibles sur le site de GEO permettaient de réanalyser une expérience donnée.
J'ai contribué à ce projet à partir de mon stage de M2BBSG en janvier 2008, sous la direction de Jean Imbert et Denis Puthier. Il portait sur le développement de fonctionnalités pour cette application. J'ai poursuivi ce projet en thèse, avec davantage de développements et d'améliorations.

4.4 Développement de l'application

Nous avons mis au point une nouvelle approche de partitionnement pour extraire de manière systématique et automatisée des groupes de gènes co-exprimés à partir de centaines de jeux de données issus de GEO et appelés GSE (pour « *Gene Serie Experiment* »).
Pour cela, nous avons utilisé l'algorithme MCL (pour « *Markov CLustering* » [Enright *et al.*, 2002]) et inclus une étape de filtrage des données permettant de ne conserver que les gènes ayant une réelle variation au sein d'une expérience et de retirer ainsi le bruit inhérent à ce genre d'expérience. Cette nouvelle méthode d'analyse a été nommée DBF-MCL pour « *Density-based filtering and Markov CLustering* » (Figure 4.2).

Les signatures transcriptionnelles (TS, pour « *Transcriptional Signature* ») correspondent à des groupes de gènes ayant des profils similaires au sein d'une expérience. Par cette stratégie innovante, nous avons extrait 18 250 TS à partir de 1 484 GSE provenant de 70 plateformes (ou GPL) de type Affymetrix™de la base de données GEO. Ces expériences correspondent à des études menées chez l'homme, le rat ou la souris, et pour lesquelles le nombre d'échantillons est supérieur à 10.

Ces 18 250 TS ont été dans un second temps annotées en utilisant la base de données de DAVID (« Database for Annotation, Visualization and Integrated Discovery », version 2005). Un enrichissement fonctionnel avec une p-valeur inférieure à 0.05 avec un test exact de Fisher corrigé par la correction de Benjamini and Hochberg a été mis en évidence pour 84% des TS.

Toutes les informations concernant l'annotation des plateformes Affymetrix, les expé-

riences, les signatures et leur annotation, ont été stockées dans une base de données relationnelle MySQL 5.0. Les données d'expression des TS sont conservées dans des fichiers plats indexés.

Nous avons développé une application java modulaire et évolutive, TBrowser, sous la forme d'un client Java distribué par Java Web Start, permettant de consulter les informations contenues dans la base de données. Un fichier jar exécutable est également proposé au téléchargement sur le site FTP du TAGC.

Cet outil est composé d'une interface graphique permettant des requêtes booléennes, utilisant donc des opérateurs logiques. Les requêtes sont de la forme : « gene1 | (gene2 & gene3) », où « | » et « & » signifient respectivement « OU » et « ET », les parenthèses servant à structurer les priorités. Il est ainsi possible d'accéder aux signatures transcriptionnelles contenant des marqueurs du cancer du sein comme : ESR1, GATA3 et FOXA1 par une requête telle que : « ESR1 & GATA3 & FOXA1 ».

Il est également possible d'exclure des gènes par l'utilisation du caractère « ! ». Ceci permet de filtrer les signatures obtenues, comme par exemple lors de requêtes avec des marqueurs de cellules T pour exclure les signatures contenant des gènes spécifiques d'un autre type de cellules (monocytes) telles que : « CD3E & CD3D & !CD14 ».

Mais les requêtes ne s'effectuent pas uniquement sur les gènes, il est en effet également possible d'interroger la base de données par annotation, sonde, plateforme, expérience. On peut ainsi récupérer toutes les signatures enrichies fonctionnellement en gènes du cycle cellulaires ("CELL CYCLE"[5,12,18]).

Les résultats d'une requête se présentent sous la forme d'une liste de TS, en relation avec les plateformes et expériences correspondantes, toutes ces informations étant chargées lors de l'affichage des résultats.

Diverses fonctionnalités ont pu être ajoutées via le développement de modules (ou plugins) afin d'utiliser les résultats des requêtes. Lors de la publication de TBrowser ([Lopez *et al.*, 2008]), seuls 3 plugins étaient présentés et mis à disposition :

- Heatmap qui permet la visualisation signature par signature des matrices d'expression (non stockées dans la base de données mais dans des fichiers plats indexés) et de leur annotation, avec toutes les informations sur les échantillons et les sondes correspondantes ;
- TBCommonGenes qui combine les listes de gènes provenant d'un groupe de signatures, permettant ainsi de définir les gènes les plus fréquemment retrouvés coexprimés dans ce groupe de signatures. Ces signatures proviennent généralement de requêtes booléennes telles que : « ESR1 & FOXA1 » pour obtenir toutes les signatures comportant au minimum ces deux gènes ;
- TBMap qui synthétise le contenu de toutes les signatures d'une plateforme de puce à ADN donnée. Afin d'obtenir une vue générale des gènes fréquemment associés ensemble au sein des signatures, des cartes transcriptionnelles ont été générées à partir des sondes des plateformes GPL96, GPL570, GPL81, respectivement les 2 plateformes humaines et la plateforme murine les plus utilisées. Ces cartes transcriptionnelles correspondent

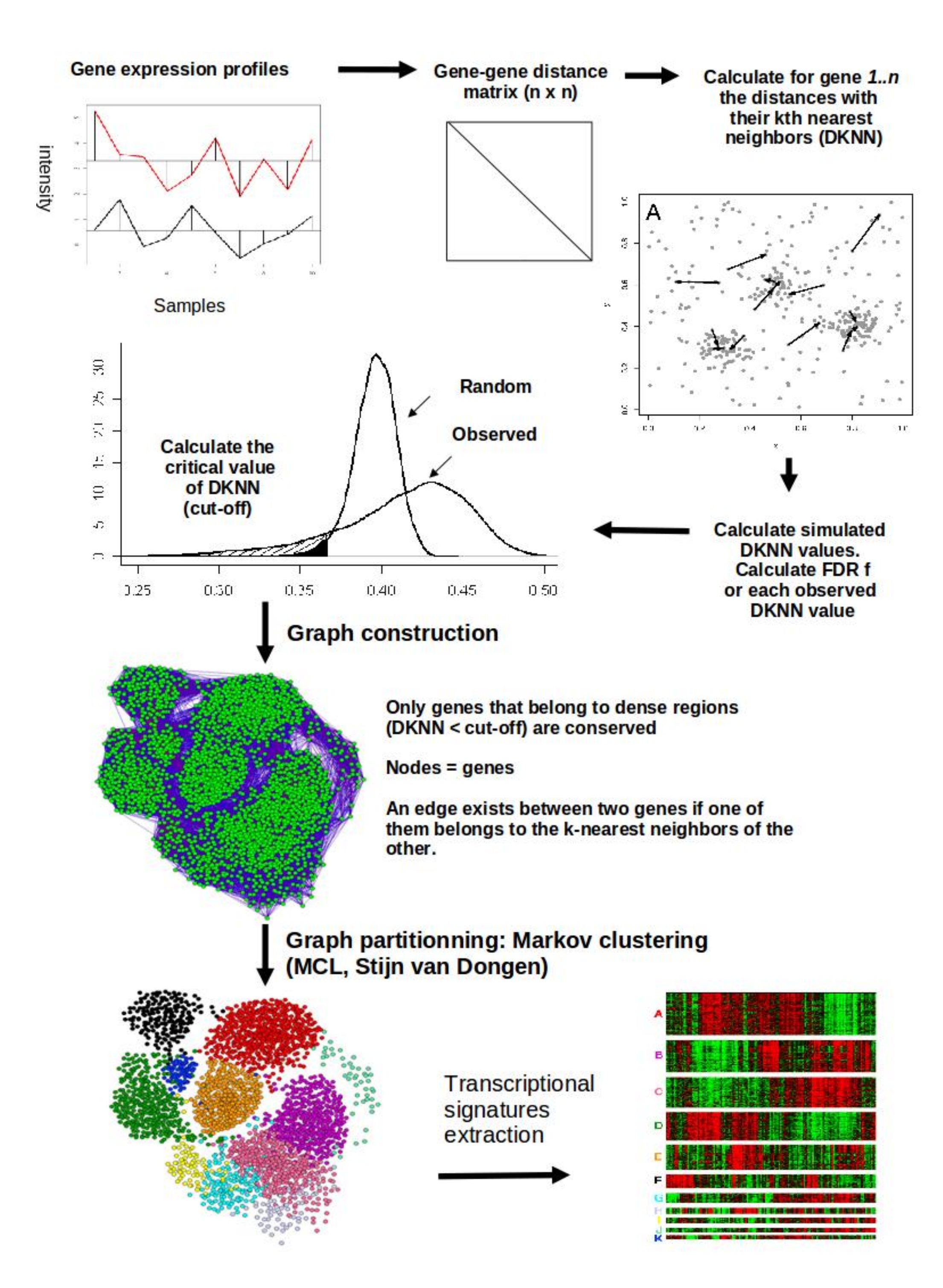


FIGURE 4.2 – Principe de l'algorithme DBF-MCL. La matrice de distance gène-gène est générée pour chaque couple de gènes. Puis elle est utilisée afin d'obtenir les distances aux k plus proches voisins (où k=150). Ces distances sont ensuite comparées à celles d'une distribution théorique, obtenue par ré-échantillonnage des distances aux k plus proches voisins observées. Ceci permet de calculer une valeur seuil, pour un FDR de 10%, par exemple. Un graphe est ensuite généré : un noeud correspond à un gène, et un arc relies un gène à ses k plus proches voisins. Enfin, l'algorithme MCL (Markov Clustering) est utilisé pour partitionner ce graphe en groupes de gènes, correspondant à des signatures transcriptionnelles.

à des matrices booléennes avec en ligne toutes les sondes d'une plateforme de puce à ADN donnée et en colonne les signatures transcriptionnelles contenues dans TBrowser et obtenues par l'algorithme DBF-MCL. Cette matrice contiendra pour une sonde i et une signature transcriptionnelle j, un « 1 » si cette signature contient cette sonde, et « 0 » si cette sonde est absente. Cette matrice subit ensuite une classification sur les lignes et les colonnes, en utilisant une distance de corrélation de Pearson.

A l'aide du plugin TBMap, on peut visualiser les cartes ainsi générées. En rouge sont représentées les sondes présentes dans une signature donnée, et en noir celles qui sont absentes. Celles-ci, une fois annotées, nous permettent d'observer le regroupement des gènes en fonction des processus biologiques et d'identifier de nouveaux gènes candidats.

Comme preuve de concept, TBrowser a été utilisé dans des études de cancers du sein (GSE1456) pour identifier des gènes spécifiques de tumeurs malignes du sein. L'utilisation de l'algorithme DBF-MCL a permis l'extraction d'une dizaine de TS, spécifiquement enrichis en gènes du cycle cellulaire, de l'adhésion cellulaire, de l'immunité (voir Lopez et al. 2008, ci-après).

Ces résultats ont été publiés dans PLoS ONE en décembre 2008.



TranscriptomeBrowser: A Powerful and Flexible Toolbox to Explore Productively the Transcriptional Landscape of the Gene Expression Omnibus Database

Fabrice Lopez[1,2]⁹, Julien Textoris[1,2,5]⁹, Aurélie Bergon[1,2], Gilles Didier[2,3], Elisabeth Remy[2,3], Samuel Granjeaud[1,2], Jean Imbert[1,2], Catherine Nguyen[1,2], Denis Puthier[1,4]*

1 Inserm U928, TAGC, Parc Scientifique de Luminy, Marseille, France, **2** Université de la Méditerranée, Marseille, France, **3** Institut de Mathématiques de Luminy, Campus de Luminy, Marseille, France, **4** ESIL, Université de Provence et de la Méditerranée, Marseille, France, **5** Service d'Anesthésie et de Réanimation, hôpital Nord - Assistance Publique, Hôpitaux de Marseille, Marseille, France

Abstract

Background: As public microarray repositories are constantly growing, we are facing the challenge of designing strategies to provide productive access to the available data.

Methodology: We used a modified version of the Markov clustering algorithm to systematically extract clusters of co-regulated genes from hundreds of microarray datasets stored in the Gene Expression Omnibus database (n = 1,484). This approach led to the definition of 18,250 transcriptional signatures (TS) that were tested for functional enrichment using the DAVID knowledgebase. Over-representation of functional terms was found in a large proportion of these TS (84%). We developed a JAVA application, TBrowser that comes with an open plug-in architecture and whose interface implements a highly sophisticated search engine supporting several Boolean operators (http://tagc.univ-mrs.fr/tbrowser/). User can search and analyze TS containing a list of identifiers (gene symbols or AffyIDs) or associated with a set of functional terms.

Conclusions/Significance: As proof of principle, TBrowser was used to define breast cancer cell specific genes and to detect chromosomal abnormalities in tumors. Finally, taking advantage of our large collection of transcriptional signatures, we constructed a comprehensive map that summarizes gene-gene co-regulations observed through all the experiments performed on HGU133A Affymetrix platform. We provide evidences that this map can extend our knowledge of cellular signaling pathways.

Citation: Lopez F, Textoris J, Bergon A, Didier G, Remy E, et al. (2008) TranscriptomeBrowser: A Powerful and Flexible Toolbox to Explore Productively the Transcriptional Landscape of the Gene Expression Omnibus Database. PLoS ONE 3(12): e4001. doi:10.1371/journal.pone.0004001

Editor: Pamela A. Silver, Harvard Medical School, United States of America

Received July 8, 2008; **Accepted** November 25, 2008; **Published** December 23, 2008

Funding: This work was supported by the Institut National de la Santé et de la Recherche Médicale (Inserm), the Canceropçle PACA and Marseille-Nice Genopole®. Fabrice Lopez was supported by a fellowship from the EU STREP grant Diamonds and through funding from the IntegraTCell project(ANR, National Research Agency). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: puthier@tagc.univ-mrs.fr

⁹ These authors contributed equally to this work.

Introduction

Microarray technology provides biologists with a powerful approach for comprehensive analyzes of cells or tissues at the transcriptional level. DNA chips are now widely used to assess the expression levels from all genes of a given organism. These data, most generally deposited in MIAME-compliant public databases, constitute an unprecedented source of knowledge for biologists [1]. As an example, until now, the Gene Expression Omnibus repository (GEO) host approximately 8,000 experiments encompassing about 200,000 biological samples analyzed using various high through-put technologies [2]. Consequently, this represents billions of measurements that reflect the biological states of cells or tissues recorded in physiological or pathological conditions or in response to various chemical compounds and/or natural molecules. As public repositories are continually expanding, we are facing the new challenge of designing new strategies to provide efficient and productive access to the available data.

To date, at least two major solutions have emerged. The first one applies a "*gene-centered perspective*", as developed in the "GEO profile" or "SOURCE" web interfaces [3]. This approach allows users to retrieve the expression profiles of a given gene in numerous curated experiments. Once a profile is selected, a list of similar profiles (*i.e.* neighbors) can be retrieved. Although GEO proposes several tools to refine the queries, cross-analysis through multiple experiments can not be performed. The second solution involves an "*experiment-centered perspective*" as developed in the "GEO DataSets" and "ArrayExpress" web interfaces [4]. This approach provides to biologists a set of classification tools to re-analyze selected experiments. Depending on the interface, supervised or unsupervised analysis (see below) can be pre-calculated or computed on demand. Again, as no meta-analysis tool is available, mining and compiling even few GEO Serie Experiments (GSE) remains a difficult and time-consuming task.

We therefore lack efficient tools allowing productive data mining of microarray databases. For example, querying whole

public microarray data using a single gene identifier is an ambiguous procedure to extract relevant co-regulated genes. Indeed, depending of the biological context, genes can be involved in different signaling pathways and may be associated with different neighbors. As a consequence, combined queries should be more appropriate to build relevant gene networks. Moreover, numerous uninformative genes exist in microarray experiments. They correspond most generally to those with low standard deviation that are outside any natural gene cluster. These genes should be discarded from analysis as they are inevitably associated with false positive neighbors. These considerations motivated the present work and the development of a new approach that follows a "*transcriptional signature centered perspective*". The goal was to build an application that would interact with a large database of transcriptional signatures and would implement efficient tools to analyze and visualize the results.

The first issue resided in the construction of a database containing high quality transcriptional signatures obtained in an automated fashion. Both supervised and unsupervised classification algorithms can be used in microarray data analysis [5]. Supervised methods aim at finding a set of genes whose expression profiles best correlate with a known phenotype. They provide a way to select informative genes by choosing the top k genes according to the results of a statistical test (*e.g.* Student's *t*-test, Significance Analysis of Microarrays, Signal to Noise Ratio, ANOVA) and by controlling the false discovery rate (FDR). In contrast, unsupervised classification approaches, achieve clustering of genes based on their respective expression profiles but are not intended to filter out uninformative genes. Some popular approaches in microarray analysis use either agglomerative methods (hierarchical clustering), partitioning methods (k-medoids, k-means, PAM, SOM, etc.) or methods aimed at capturing informative dimensions (PCA). A filtering step is most generally applied prior to unsupervised classification. One can select genes with high standard deviations, those displaying a proportion of values above a user-defined threshold or those having a given maximum (or minimum) value. However this procedure is extremely subjective and the number of selected genes may be over or under estimated. Finally, another limit of classical unsupervised methods also resides in their inability to accurately identify the actual number of clusters if no further argument is provided to the algorithm. As a consequence, additional algorithms for unsupervised classification have been proposed such as Quality Cluster algorithm (QT_Clust) [6], CHAMELEON [7] or Markov CLustering (MCL) [8]. However, none of them address both the filtering and partitioning issues. MCL is a graph partitioning algorithm whose ability to solve complex classification problems has been underlined in many applications including protein-protein interaction networks [9], sequence analysis (TRIBE-MCL) [10] or microarray analysis (geneMCL) [11]. In a graph representation of microarray data, nodes stand for genes and edges represent profile similarities between genes. As processing the full graph for partitioning is time-consuming and computer-intensive the geneMCL algorithm has to be run on a subset of genes that are selected using classical filters (*e.g.* high standard deviation or fold-change). As such a filtering procedure is not well suited for automated analysis of numerous experiments; we developed an adaptive density-based filter (DBF) whose goal is to isolate automatically informative genes from a dataset. Selected genes are next used to construct a graph that is subsequently partitioned using MCL. This modified version of MCL algorithm was termed DBF-MCL for "Density Based Filtering and Markov CLustering".

In the present paper, we show that DBF-MCL provides very good results both on simulated and real datasets. The algorithm was run on 1,484 microarrays datasets (46,564 biological samples) performed on various Affymetrix platforms (human, mouse and rat). This led to the identification of 18,250 transcriptional signatures (TS) whose corresponding gene lists were tested for an enrichment in terms derived from numerous ontologies or curated databases using the DAVID knowledgebase [12] (Gene Ontology, KEGG, BioCarta, Swiss-Prot, BBID, SMART, NIH Genetic Association DB, COG/KOG, etc.) (see Figure S1 for an overview of the data processing pipeline). Informations related to biological samples, experiments, TS composition, TS associated expression values and TS keyword enrichment scores were stored in a relational database. A Java application, TBrowser (TranscriptomeBrowser), was developed and deployed using Java Web Start technology. Combined queries that can be done with an extended set of Boolean operators allow user to rapidly select sets of TS containing (or not) a given list of gene symbols. Based on these TS, a list of frequently observed neighbors can be created. As each TS is linked to a set of biological keywords (derived from ontologies), user can also search for those enriched in genes involved in specific biological processes. We show that TBrowser can be used to mine productively hundreds of experiments and to reveal underlying gene networks. Furthermore, using this unprecedented collection of TS we built the first synthetic transcriptional map of all human microarray data performed on Affymetrix HG-U133A platform and currently available in the GEO database.

Results

DBF-MCL algorithm

Conventional algorithms used for unsupervised classifications of gene expression profiles suffer from two main limitations. First, they do not filter out uninformative profiles and second, they are not able to find out the actual number of natural clusters in a microarray dataset. We can considerer genes as points located in a hyperspace whose number of axes would be equal to the number of biological samples. As it is difficult to perceive high-dimensional spaces, a common way to illustrate classification methods is to use a 2D representation. In Supplemental Figure S2, each point represents a gene and we are interested in isolating dense regions as they are populated with genes that display weak distances to their nearest neighbors (*i.e.* strong profile similarities). To isolate these regions we can compute, for each gene, the distance with its k_{th} nearest neighbor (DKNN). If k is relatively small, DKNN should be smaller for all genes falling in a dense area. Thus, the filtering procedure used in DBF-MCL starts by computing a gene-gene distance matrix D. Then, for each gene, DBF-MCL computes its associated DKNN value (with k being set typically to 100 for microarrays containing 10 to 50k elements). Distributions of DKNN values observed with both an artificial and a real dataset (Complex9RN200 and GSE1456 respectively, see thereafter for a description) are shown in Figure S3A and S3B (solid curve). The asymmetrical shape of the distribution observed in Figure S3B suggests the presence of a particular structure within the GSE1456 microarray dataset. Indeed, the long tail that corresponds to low DKNN values could indicate the existence of dense regions. The fact that regions of heterogeneous densities exist in the Complex9RN200 artificial dataset is even clearer as a bimodal distribution is observed. Next, we would like to define a critical DKNN value below which a gene can be considered as belonging to a dense area and that would depend on the intrinsic structure of the dataset. To this end, DBF-MCL computes simulated DKNN values by using an empirical randomization procedure. Given a dataset containing n genes and p samples, a simulated DKNN value is obtained by sampling n distance values from the gene-gene distance matrix D and by extracting the k_{th}-smallest value. This procedure is repeated n times

to obtain a set of simulated DKNN values S_i. As shown in Figure S3 (dotted line), the variance of the simulated DKNN values is very low compare to that observed using the real dataset. Indeed, we can think of simulated DKNN values as the distances to the k_{th} element if no structure existed in the associated space. In this case, we would expect elements to be uniformly spread throughout the space and the variance of DKNN value to be low. In practice several sets $S_{1..q}$ are computed and thus several distributions of simulated DKNN values are obtained. For each observed DKNN value d, a false discovery rate (FDR) value is estimated by dividing the mean number of simulated DKNN below d by the number of observed value below d. The critical value of DKNN is the one for which a user-defined FDR value (typically 10%) is observed. Given a set of selected genes, the next issue is to partition them into homogeneous clusters. This step is achieved through a graph partitioning procedure. In the created graph, edges are constructed between two genes (nodes) if one of them belongs to the k nearest neighbor of the other. Edges are weighted based on the respective coefficient of correlation (*i.e.*; similarity) and the graph obtained is partitioned using the Markov CLustering Algorithm (MCL).

Performances of DBF-MCL on Complex9RN200 dataset

To test the performances of DBF-MCL algorithm we used a modified version of the complex9 dataset which was used earlier by Karypis *et al.* [7]. Since DBF-MCL is designed to handle noisy datasets, 200% of normally distributed random noise was added to the original data. The resulting dataset (which will be referred as Complex9RN200 thereafter, see Figure S4A and S4B) shows some difficulties for partitioning since it is composed of a noisy environment in which arbitrary geometric entities with various spacing have been placed. The two main parameters of DBF-MCL are k that controls the size of the neighborhood and the inflation I (range 1.1 to 5) which controls the way the underlying graph is partitioned. The effect of k on the selection of informative elements is shown in Figure S5A (Euclidean distance was used for this dataset). A steep ascending phase and a slow increasing phase (starting from a k values close to 40) were observed. This confirms the existence of areas with heterogeneous densities. In fact, the transition between the two phases reflects the transition from dense to sparse regions. Indeed, datasets produced with k values above 40 contain noisy elements (Fig S4C). In contrast, choosing k values in the ascending phase ensure the achievement of noise-free datasets. In the case of artificial data, satisfying partitioning results were obtained with inflation values close to 1.2 (Fig S4D–G) although in some cases some of the shapes were merged in a manner that appears to be meaningful (Fig S4E and S4G). We then compared DBF-MCL to several algorithms commonly used in microarray analysis. All of them were run multiple times with various parameters and the best solution was kept. In all cases, the Euclidean distance was used as a distance measure between elements. As these algorithms are not well-suited for noisy data, they were run on the 3,108 points extracted using DBF-MCL ($k = 20$). Also it is difficult to compare those algorithms to one another, some of them obviously failed to identify the shapes. Indeed, although k-means was run 10 times with random initial starts (and the right number of centers) it led to a very poor partitioning result (Fig S4J). Cluster Affinity Search Technique (CAST, Fig S4K) and Quality Cluster algorithm QT_CLUST (Fig S4I), gave also poor results as did the Self-Organizing Map (SOM) (data not shown). Hierarchical clustering was run with single linkage as arguments and the obtained dendrogram was then split into 9 clusters (Fig S4H). Patterns were well recognized using this method but prior knowledge of the number of clusters is a prerequisite. Thus both DBF-MCL and hierarchical clustering are able to properly identify complex shapes in a 2D space. The main benefit of using DBF-MCL resides in its ability to extract relevant informations from a noisy environment. However, a range of optimal values for inflation parameter needs to be defined to get the best results.

Performances of DBF-MCL on GSE1456 dataset

Next, DBF-MCL was tested with microarray data to explore its effectiveness in finding clusters of co-regulated genes. To this end, we used the microarray data from Pawitan *et al.* [13], who studied gene expression profiles in a large cohort of Swedish patients affected by breast cancer. This experiment is recorded as GSE1456 in the GEO database. All sample ($n = 159$) have been hybridized onto the GPL96 platform (Affymetrix GeneChip Human Genome U133 Array Set, HG-U133A). The complete dataset (22,283 genes) was used for analysis. Figure S5B, shows the number of informative genes obtained with various k values. Again, two phases were observed suggesting that regions with heterogeneous densities exist in the GSE1456 dataset. As expected, the transition from dense to sparse regions was less marked than in the artificial dataset. A k value of 100 was chosen to allow the extraction of a large part of data that can be considered as noise-free. This value led to the selection of 4,470 elements out of the whole dataset (Fig. 1A–B). The graph partitioning procedure, using default MCL parameters ($I = 2$), generated 11 highly homogeneous clusters (Fig. 1C–F). As with the Complex9RN200 dataset, the results were very consistent with those obtained using hierarchical clustering although for some genes the clustering results differed (Fig. 1E). Importantly, partition results were not very sensitive to inflation values. Indeed, 10 and 12 clusters were observed with I set to 1.5 and 2.5 respectively (data not shown). All signatures were then submitted to functional enrichment analysis. A summary of the results is given in Figure 1G. As expected for a breast cancer dataset, TS were found to be related to (i) immune response (T-lymphocyte activation, B-lymphocyte activation and interferon alpha), (ii) primary metabolism (cell cycle, ribosome biogenesis, nuclear phosphorylation and transcription) which is probably reminiscent of tumor aggressiveness (iii), modification of local environment (extracellular matrix and cell adhesion) which could sign metastasis potential of each sample, (iv) and estrogen receptor status of breast tumors (estrogen response pathway). Altogether, these results underline the ability of DBF-MCL algorithm to find natural gene clusters within a randomly selected dataset. Indeed, for numerous additional microarray datasets hierarchical clustering results and DBF-MCL results were compared. As illustrated in Figure S5B for a representative set of experiments, setting k to 100 allows in all cases to delete noisy elements and to select only informative genes in a microarray dataset. Interestingly, in all cases meaningful partitioning results were obtained using inflation parameter set to 2.

Systematic extraction of TS

We next applied DBF-MCL algorithm to all experiments performed on human, mouse and rat Affymetrix microarrays and available in the GEO database (33 platforms, Supplementary Table S1 and S2). Only experiments containing more than 10 biological samples were kept for analysis. Overall, this dataset includes 46,564 biological samples hybridized in the context of 1,484 experiments. Each experiment was analyzed independently and subjected to TS discovery process ($k = 100$, $FDR = 10\%$, $S_{1..3}$, $Inflation = 2$). As mentioned in the Material and Methods section, we rank-transformed data from each biological sample to get a common input for DBF-MCL algorithm and to allow analysis of a large broad of experiments whose normalization status is frequently unknown. Furthermore, a distance based on Spear-

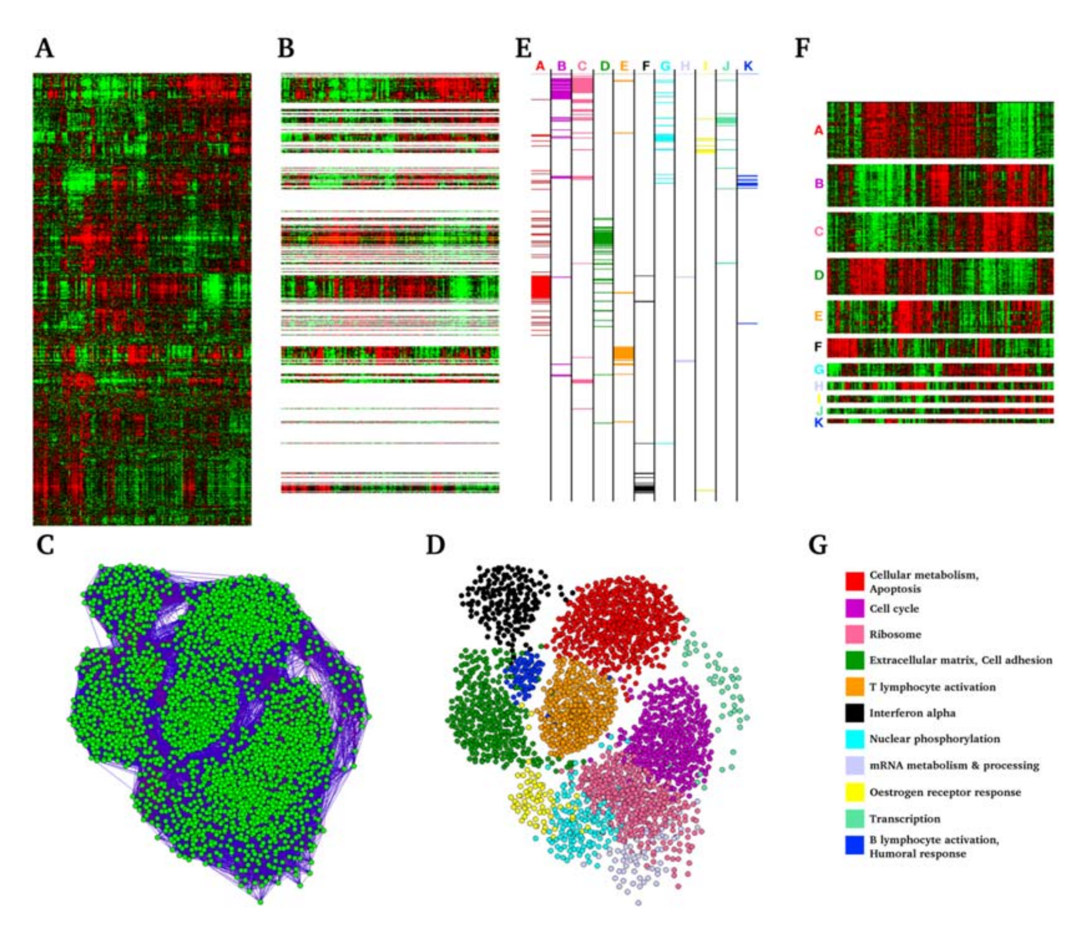


Figure 1. Results obtained with the GSE1456 dataset. DBF-MCL was run with GSE1456 as input (k = 100, FDR = 10%, $S_{1..3}$, Inflation = 2). **(A)** Hierarchical clustering of the GSE1456 dataset. **(B)** Same as (A) but only informative genes are displayed. **(C)** The graph constructed with the 4,470 selected genes. **(D)** The graph after MCL partitioning. Each point is colored according to its associated class. **(E)** Correspondence between hierarchical clustering and DBF-MCL results. **(F)** TS obtained for GSE1456 **(G)** Functional enrichment associated with these TS.
doi:10.1371/journal.pone.0004001.g001

man's rank correlation coefficient was used for k_{th}-nearest neighbor computation. This rank-based distance is known to be clearly more resistant to outlying data points than Pearson-based distance and thus ensured the selection of genes belonging to unmistakable clusters. The full pipeline was run on a server equipped with 6 CPUs and took about 4 days to complete. For the sake of clarity, only results obtained with GPL96 which is the most widely used Affymetrix microarray platform will be presented in this section (311 experiments related to GPL96 were analyzed, 12,752 hybridized samples). On average, 4,341 probes (min = 832, max = 5,849) per expression matrix were declared as informative by DBF-MCL suggesting that routinely 20% of the 22,283 probes measured on the HG-U133A array belong to a natural cluster. Graph partitioning generated on average 10.8 clusters (min = 2, max = 29) for each experiment and each cluster contained approximately 400 probes corresponding in average to 370 distinct gene symbols. Figure 2 shows a summary of these results. As expected no clear correlation was observed between the number of selected genes and the number of samples in the experiments which demonstrates the robustness of the filtering process. In contrast, a trend to produce more clusters in experiments containing few samples was observed. This was notably marked in experiments containing 10 to 15 samples. Such a bias is classical in data analysis. Indeed, if numerous values (*i.e.* samples) are used to estimate the expression profile of a given gene, outliers will have weak impact on distance calculation and the gene will be assign to the expected cluster. In contrast, when only few values are available, each of them has a greater impact on distance calculation. This results in producing more clusters with some of them having centers close to one another. This bias is also presumably amplified by the fact that small sample sets contain most generally a greater biological diversity compared to large sample sets as they contain fewer replicates. Overall, our analysis of GPL96 related experiments gave rise to 3,377 TS. The full analysis on the 33 Affymetrix platforms produced 18,250 TS which correspond to 220 millions of expression values. Partitioning results where manually checked for a large panel of experiments. Although, results seemed perfectible in few cases, they always appeared to be rational.

The TBrowser interface

Comprehensive information on samples, experiments, probes and genes were stored in a mySQL relational database. A flat file indexed on TS IDs was used to store TS expression data. This solution was preferred because it turned out to be an excellent alternative to database for retrieving rapidly expression values for the selected TS. We next developed TBrowser, a Multitier architecture system composed of (i) a "heavy client" written in JAVA (presentation Tier), (ii) a servlet container (logic tier) and (iii) a back-end database (data tier). The client application allows user

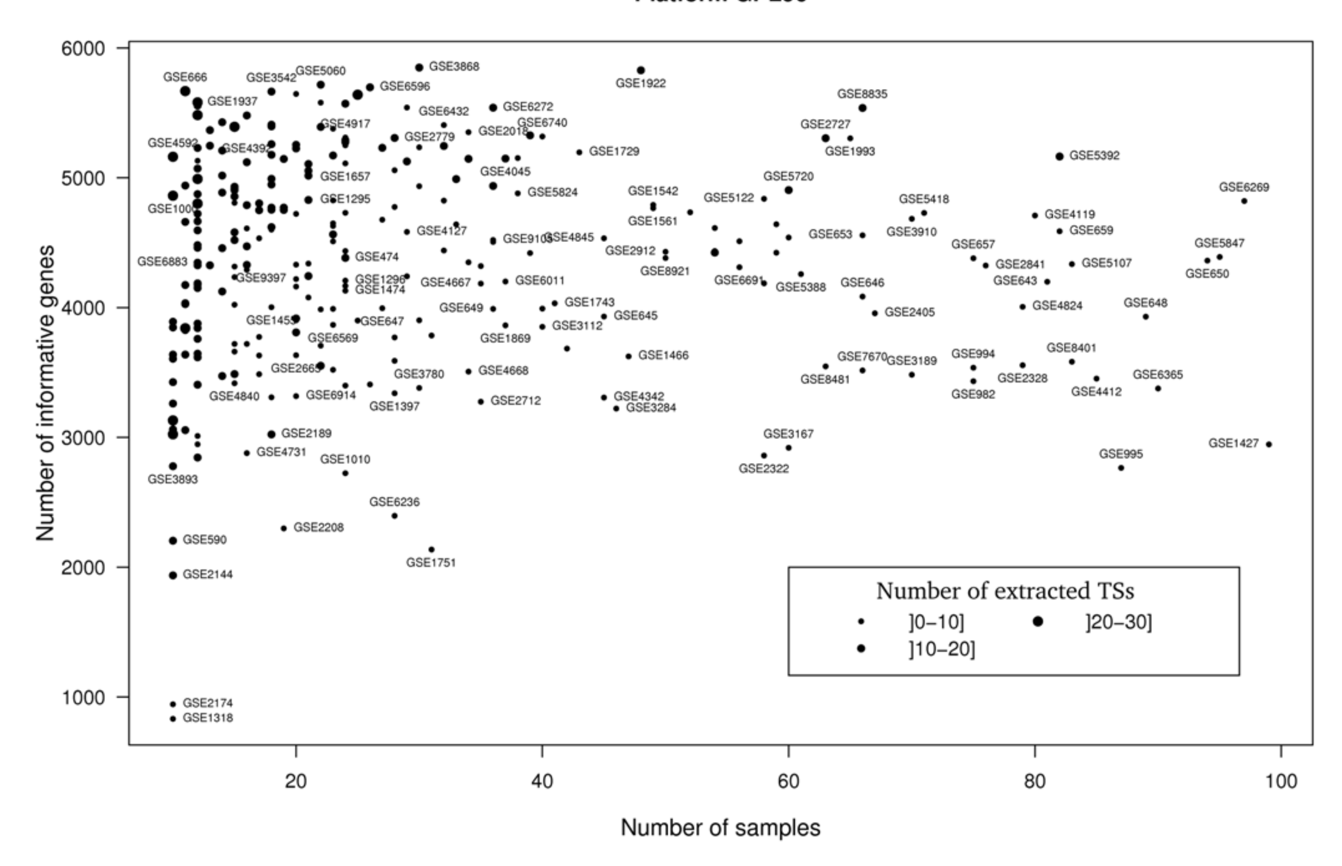


Figure 2. Large scale TS extraction from GPL96 experiments. DBF-MCL was run with default parameters (k = 100, FDR = 10%, $S_{1..3}$, Inflation = 2). X axis corresponds to the number of samples in the experiment and Y axis to the number of informative genes. For each experiment, the number of associated TS is represented by the size of the dot. For clarity purpose only experiments with less than 100 samples are represented. Furthermore, the name of only some of them is displayed.
doi:10.1371/journal.pone.0004001.g002

to query TBrowser database using six methods: by gene symbols, by probe IDs, by experiments, by microarray platform, by ontology terms (annotation) or by TS. Three of them (gene symbols, probe IDs, and annotation methods) accept a list of operators that control the way a query is to be processed. One may take advantage of these operators to create complex queries using the AND operator (&), the OR operator (|), the NOT operator (!) or using additional characters such as the quote or parenthesis (reader may refer to the user guide for additional explanations and informations). The main window of TBrowser is made of five panels (Fig. 3). The search panel is the main entry as it is used (i) to define the search method, (ii) to write the queries, (iii) to launch database interrogation and (iv) eventually to filter out some of the TS. Filters can be applied to select species of interest and to control the sizes (number of samples and number of genes) of the TS that one wants to analyze. The results area can display two panels: the list of queries the user launched during his session and the list of TS that correspond to the currently selected query. Double-clicking on one (or several) TS send it (them) to the selected plugin. The information area is used to display various informations about the selected TS whereas the plugin area is used to select one of the currently installed plugins. Finally, the plugin display panel manages the display of the currently selected plugin. To date, eight plugins have been developed (three of them are presented in this article). The Heatmap plugin is composed of two main panels: the heatmap on the left and the annotation panel on the right (Fig. 3). The Heatmap panel displays a color-coded image of TS expression values. In this representation, each row corresponds to a probe and each column to a sample. Additional informations, such as external links, can be retrieved by single-click on genes or samples. Functional enrichment informations are available on the right. The TBCommonGenes plugin was developed to compare gene composition of several TS and will be presented in the next section. Finally the TBMap plugin that can be used to visualize a summary of transcriptional regulation events observed in a given microarray platform will be presented in the last paragraph of the results section.

Meta-analysis of public microarray data using TBrowser: a case study

TBrowser can be used in many biological contexts to point out relevant experiments and construct robust gene networks. Several peer-reviewed publications have highlighted the joint regulation of the estrogen receptor-a (ESR1/ER-α), GATA3 and FOXA1 in breast cancer cells [14]. Although some of these reports have associated entry in the GEO database, retrieving neighbors of GATA3, FOXA1 and ESR1 remains a time consuming and difficult task using existing tools. As a consequence, these informations are reserved to those with strong bioinformatics skills although they are of primary interest to the biologist. Using the TBrowser search engine, this task can be translated into a very simple Boolean query, ''ESR1 & GATA3 & FOXA1'', which will

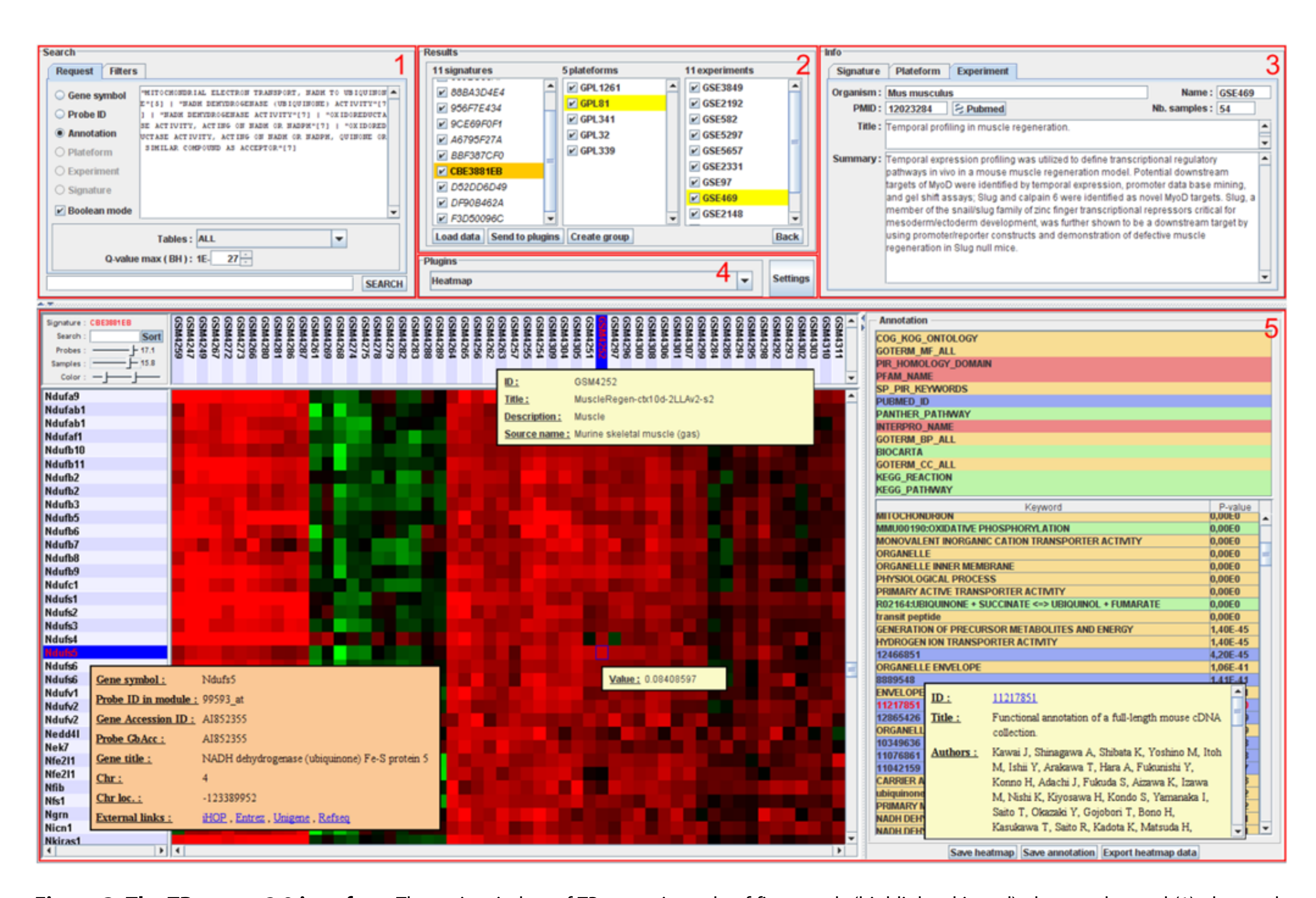


Figure 3. The TBrowser 2.0 interface. The main window of TBrowser is made of five panels (highlighted in red): the search panel (1), the results panel (2), the information panel (3), the plugins panel (4) and the plugin display panel (5). This example shows the expression profiles of genes contained in the TS CBE3881EB derived from GSE469 ("Temporal profiling in muscle regeneration"). Annotation panel shows that this TS is highly enriched in genes related to ATP synthesis.
doi:10.1371/journal.pone.0004001.g003

be almost instantaneously proceeded by the server. With the current database release, this produces a list of 16 TS (see Table 1) containing on average 508 probes (range: 82–1,572) and which were obtained using various microarray platforms (GPL96, GPL570, GPL91). Interestingly, all these TS are related to experiments performed on breast cancer cells underlying the high specificity of this gene list (Table 1). The TBCommonGenes plugin indicates that in addition to ESR1, GATA3 and FOXA1 two genes (ANXA9 and ERBB4) are found in all 16 TS. Importantly, 63 genes are found in at least 10 out of the 16 selected TS (63%). As expected, this list contains numerous markers of breast cancer cells whose expression specificity was previously reported by other (notably ERBB3, XBP1, KRT18, IL6ST, CREB1, TFF1, TFF3; see Supplementary Table S3). Thus TBrowser can be used to perform meta-analysis of microarray data in a platform-independent manner providing high confidence gene lists. However, one can also focus the analysis on a unique platform. Indeed, the transcriptional signatures 3DE64836D, B79B1C0B9 and E2E620F40 that were derived from the GPL570 platform (which measures over 47,000 transcripts) share a list of 68 genes. Many of them correspond to poorly characterized genes (for example, C17orf28 C1orf64, KIAA1370, KIAA1467, LOC143381, LOC400451, LOC92497 and ZNF703). This example clearly demonstrates the superiority of TBrowser over conventional approaches as it can be used, easily and productively, to create robust sets of transcriptionally related genes whose subsequent analysis may be crucial in defining new therapeutic targets.

Using annotation terms to mine public microarray data

Based on the systematic functional enrichment analysis, the vast majority of TS (84%) have a set of associated biological terms (only functional enrichment with q-value<0.01 are stored in the database). One can search for TS related to functional terms of the DAVID knowledgebase (*e.g.* "nervous system development"). More interestingly, multiple terms can be combined with Boolean operators. Searching for TS which contain genes located in the 6p21.3 and 14q32.33 chromosomal regions (major histocompatibility complex and human immunoglobulin heavy-chain locus respectively) and which contain T-cell specific genes, can be translated as: 6p21.3[4] & 14q32.33[4] & "T CELL ACTIVATION"[5,12] ([4] = cytoband term, [5] = GO term, [12] = Panther pathways term). As chromosomal aberrations do occur frequently in cancer our approach can also be used to perform systematic cytogenetic analysis. Indeed, throughout our analysis, 2,208 functional enrichments related to 360 human cytobands were observed and stored in the database. As an example, TS with very strong enrichment (q-value$<1.10^{-20}$) for any of the human cytobands stored in the database are presented in Table 2. The first one is related to atopic dermatis analysis (skin biopsies) and contained 24% of genes located in 17q12-q21. They correspond to genes encoding for the keratin and keratin-associated protein families (KRT17, KRT27, KRTAP1-5, KRTAP17-1, KRTAP3-1, KRTAP3-3, KRTAP4-10, KRTAP4-12, KRTAP4-13, KRTAP4-15, KRTAP4-2, KRTAP4-3, KRTAP4-5, KRTAP4-8, KRTAP4-9, KRTAP9-2, KRTAP9-3, KRTAP9-4 and

Table 1. Transcriptional signatures containing Affymetrix probes for ESR1, GATA3 and FOXA1.

TS ID[1]	Genes[2]	Probes[2]	Samples[2]	Sample type	GSE ID	GPL ID	Author	PubMed IDs
0F2635383	1190	1572	23	Cell lines	GSE6569	GPL96	Huang F et al 2007	17332353
3DE64836D	102	143	62	Tissue	GSE7904	GPL570	unpublished 2007	-
59A18E225	690	893	121	Both	GSE2603	GPL96	Minn AJ et al 2005	16049480
6C975B20B	88	96	26	Tissue	GSE6772	GPL96	Klein A et al 2007	17410534
6C975B290	88	96	26	Tissue	GSE6596	GPL96	Klein A et al 2007	17410534
7150E17F6	868	1032	34	Cell lines	GSE4668	GPL96	Coser KR et al 2003	14610279
8059848B4	200	250	251	Tissue	GSE3494	GPL96	Miller LD et al 2005	16141321
84E5E1077	694	883	198	Tissue	GSE7390	GPL96	Desmedt C et al 2007	17545524
8F69864F9	68	82	95	Tissue	GSE5847	GPL96	Boersma BJ et al 2007	17999412
A151D5695	297	361	58	Tissue	GSE5327	GPL96	Minn AJ et al 2007	17420468
B79B1C0B9	270	380	47	Tissue	GSE3744	GPL570	Richardson AL et al 2006	16473279
BDB6D8700	550	679	104	Tissue	GSE3726	GPL96	Chowdary D et al 2006	16436632
D8F0B528C	125	152	159	Tissue	GSE1456	GPL96	Pawitan Y et al 2005	16280042
E2E620F40	448	616	129	Tissue	GSE5460	GPL570	unpublished 2007	-
EA9669A21	219	251	158	Tissue	GSE3143	GPL91	Bild AH et al 2006	16273092
F310ACC36	519	646	49	Tissue	GSE1561	GPL96	Farmer P et al 2005	15897907

[1]Transcriptional signature ID.
[2]Total number.
doi:10.1371/journal.pone.0004001.t001

KRTAP9-8). This signature is notably annotated as being enriched in genes related to PMID 11279113 ("Characterization of a cluster of human high/ultrahigh sulfur keratin-associated protein genes embedded in the type I keratin gene domain on chromosome 17q12-21") [15] and in genes related to the PIR keyword "multigene family". Furthermore, several signatures, of Table 2 are related to melanoma and six of them were observed in the GSE7127 experiment [16]. Although data from Table 2 would deserve further analysis they are most likely related to gain or loss of genetic material in tumors. Indeed, gain of 8q is frequently observed in a number of tumor types (including melanoma and ovarian tumors) and this region is known to contain the c-myc oncogene at 8q24.21. Interestingly, in several cases, contiguous cytobands were significantly enriched suggesting a large deletion or amplification of genetic material in these tumors (TS 60E29DA83 is enriched in genes from 8q13, 8q21.11, 8q22.1, 8q22.3, 8q24.13 and 8q24.3 cytobands). In the same way, loss of genetic material of the long arm of chromosome 11 occurs in primary melanoma but is even more frequent in metastatic tumors (TS A93ED7519 is enriched in genes from 11q21, 11q23.3 and 11q24.2 cytobands). Altogether, these results underline the versatility of TBrowser and its ability to extract hidden and meaningful informations from published or unpublished microarray data. Indeed, the cytogenetic results presented in Table 2 were not discussed by the authors in the corresponding articles.

A synthetic view of all GPL96 related experiments

The paradigm that genes from a TS share functional relationships is now widely accepted and constitutes the basis of transcriptome analysis [17]. However, each of these TS is rather

Table 2. Transcriptionnal signatures displaying high enrichment (q value<1.10−20) for any of the human cytoband tested.

TS ID[1]	Enrich.[2]	Cytoband	q.value	Sample type	GSE ID	GPL ID	Authors	PubMed ID
3DA3C8345	24%	17q12-q21	$1.7.10^{-39}$	Skin	GSE5667	GPL97	Plager DA et al 2007	17181634
43CC3EF57	9%	8q24.3	$7.0.10^{-32}$	Melanoma	GSE7153	GPL570	Unpublished 2007	-
60E29DA83	16%	8q24.3	$6.8.10^{-24}$	Melanoma	GSE7127	GPL570	Johansson P et al 2007	17516929
60E581184	26%	17q25.1	$5.5.10^{-23}$	Melanoma	GSE7127	GPL570	Johansson P et al 2007	17516929
60E6B4129	35%	20p13	$1.6.10^{-26}$	Melanoma	GSE7127	GPL570	Johansson P et al 2007	17516929
60E96FF1E	28%	6p21.3	$1.2.10^{-28}$	Melanoma	GSE7127	GPL570	Johansson P et al 2007	17516929
60EC95F6A	17%	7q22.1	$6.3.10^{-31}$	Melanoma	GSE7127	GPL570	Johansson P et al 2007	17516929
60EEBD669	32%	11q23.3	$1.4.10^{-26}$	Melanoma	GSE7127	GPL570	Johansson P et al 2007	17516929
B4C95CF18	42%	8q24.3	$1.1.10^{-36}$	Ovary	GSE6008	GPL96	Hendrix ND et al 2006	16452189
A93ED6519	16%	11q23.3	$6.9.10^{-23}$	Melanoma	GSE7152	GPL570	Packer LM et al 2007	17450523
A93DB01ED	11%	7q22.1	$9.5.10^{-30}$	Melanoma	GSE7152	GPL570	Packer LM et al 2007	17450523

[1]Transcriptional signature ID.
[2]Enrichment: Proportion of non redondant genes from the TS that are located in the corresponding cytoband.
doi:10.1371/journal.pone.0004001.t002

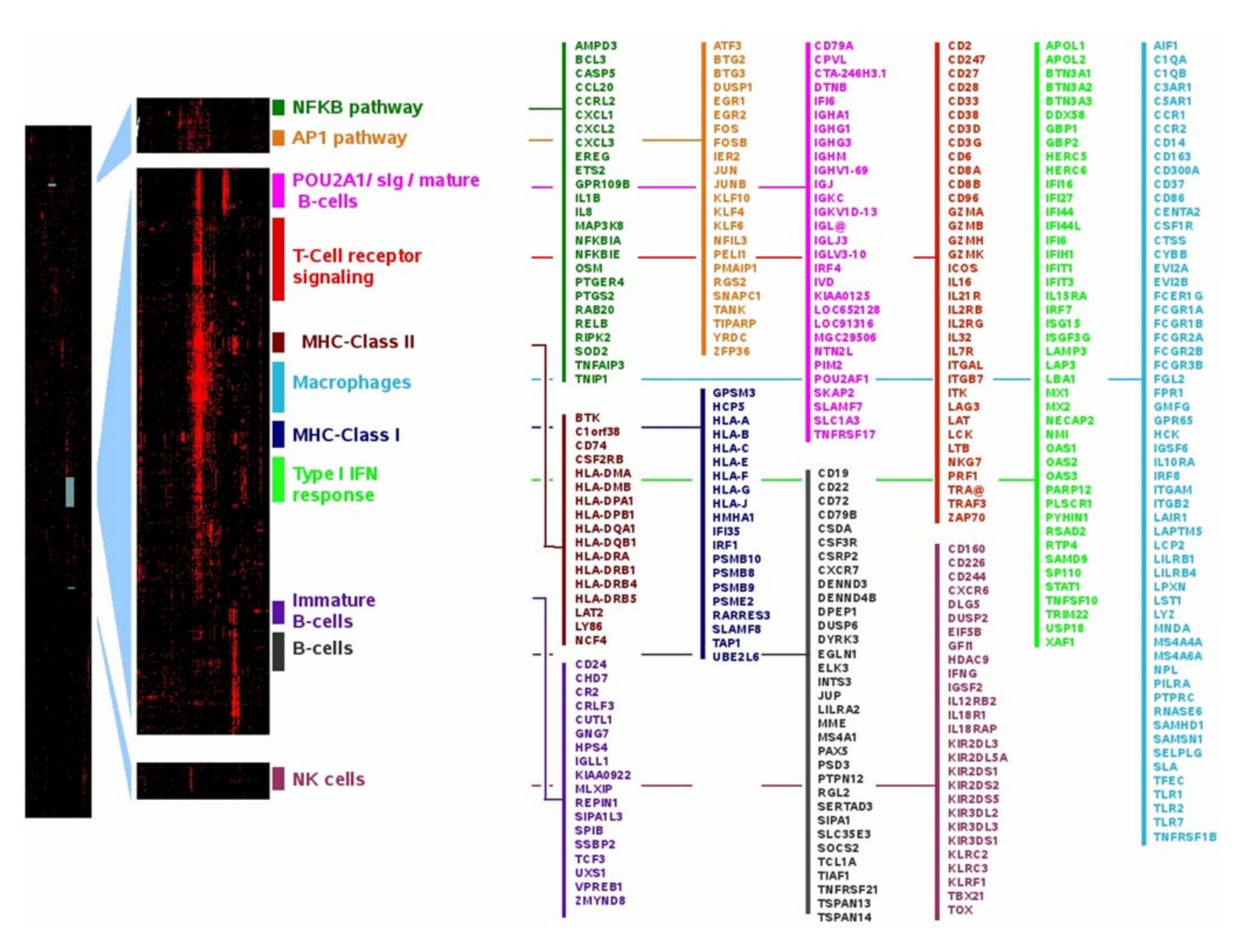


Figure 4. The transcriptional MAP associated with GPL96 related experiments. (A) A low resolution image made of 22,215 probes from GPL96 platform as rows and 3,114 GPL96 specific TS as columns. Red color indicates the presence of a gene in the corresponding TS (default to black). **(B)** Zooms of the corresponding areas showing some immune system related meta-signatures. **(C)** Representative genes that fall into these clusters.
doi:10.1371/journal.pone.0004001.g004

associated to multiple underlying pathways whose components and limits are unclear. Our difficulty in depicting comprehensive maps for pathways is illustrated by existing discrepancies, for instance, between those proposed by BioCarta, KEGG and GeneMAPP. We reasoned that the more frequently two genes fall in the same TS, the more likely these genes belong to the same core functional network. To test this hypothesis, we produced a Boolean matrix with 22,215 probes from GPL96 platform as rows and 3,114 GPL96 specific TS as columns (only TS containing 30 to 1500 probes were included). This matrix was filled with zero and elements were set to 1 if a given gene was observed in the corresponding TS. Hierarchical clustering with uncentered Pearson's correlation coefficient was used to reveal genes frequently associated to the same TS. Given the order of the resulting matrix, it could not be visualized on a desktop computer using conventional software (*i.e.*; Treeview, MeV). We thus developed the TBMap plugin which allows one to visualize the map but also to superimpose a user-defined or a KEGG-related gene list. As expected, most of the clusters where obviously enriched in genes involved in similar biological processes (Protein biosynthesis/Ribosome function, oxidative phosphorylation, cell cycle, fatty acid metabolism, valine leucine and isoleucine degradation, extracellular matrix, breast cancer cells, structural constituent of muscles, neuronal processes, etc.). This was particularly clear when KEGG pathway informations were superimposed (see Figure S6). The Figure 4 presents some of the clusters that were identified as related to immune system functions. We could find a signature defining T cells that contained numerous cell-surface markers (e.g. TCA@, CD2, CD3G, CD6, IL2RB, IL2RG, IL7R, IL21R and ICOS), signaling genes (ZAP70, LAT, LCK, ITK) and cytotoxicity-related genes (GZMA, GZMB, GZMH, GZMK and PRF1). Concerning B-cells, three clusters were observed. A large signature contains mature B-cell markers (CD19, CD22, CD72 and CD79B) and transcription factors important in B-cell development such as PAX5 and TCL1A. A second signature contains POU2AF1/OBF-1, together with its described targets: genes coding for immunoglobulin (IGHG1, IGHG3, IGHA1, IGHM, IGJ, IGKC and IGL) and the B-cell maturation factor, TNFRSF17/BCMA [18,19]. The third B-cell signature contains cell surface markers found in immature B-cells (CD24, VPREB1, IGLL1/CD179B and CR2/CD21) in addition to transcription factors known to play a crucial role during early B-cell development (TCF3, SPIB and CUTL1). The NK signature contains eight genes of the Killer cell immunoglobulin-like receptors (KIR) family, 3 genes of the killer cell lectin-like receptor family in addition to other markers whose expression has

been reported on the surface of NK cells (CD160, CD244/2B4 and CD226) [20,21,22]. It also contains TBX21/T-bet together with IL18R1, IL18RAP, IL12RB2 and IFNG. Importantly, the IL12/IL18 combination has been shown to be potent inducers of both TBX21/T-bet and IFNG in NK cells[23,24]. In addition to MHC-Class I, MHC-Class II and macrophage related signatures, two pathways related to immune function are presented in Figure 4. The AP1 pathway is made of the prototypical immediate early genes and contains numerous transcription factors (EGR1, EGR2, FOS, FOSB, IER2, JUN, JUNB, KLF6, KLF4, KLF10, ATF3, BTG2 and BTG3) whose complex interplay has been reported earlier. Finally, a NFKB signature was also observed which, again, contains prototypical regulators (NFKIA, NFKIE, RELB, BCL-3 and MAP3K8/TPL2) and known targets (CCL20, CXCL3, IL1B, IL8 and SOD2). Altogether, these results underline the high relevance of the signatures obtained using this compilation of TS derived from GPL96 related GEO experiments.

Discussion

In the present paper, we present the construction of a unique collection of TS that summarize almost all human, mouse and rat Affymetrix microarray data stored in the GEO database. TBrowser constitutes a highly powerful search engine that makes it possible to perform easily platform independent meta-analysis of microarray data. This can be considered as a real improvement over classical approaches and softwares as it provides easy and productive access to data without the need of any programming skills. Indeed the simple use of an extended set of operators proved to be sufficient to construct robust gene networks and assign poorly characterized genes to relevant biological pathways. As a consequence, it is particularly well suited to compare results obtained through microarray, ChIP-on-chip, ChIP-seq, CGH or protein-protein interaction experiments to those previously stored in the GEO database.

In all tested experiments, we found that DBF-MCL gives very good results both on simulated datasets and real microarray datasets. Although Lattimore *et al* proposed another MCL-based algorithm (geneMCL) we were unable to compare our results with their implementation as the software is no longer available nor maintained. However, DBF-MCL was run on the full van't Veer DataSet [25] (117 biological samples) that was used by Lattimore and collaborators in the original paper. In their report, the authors used a subset of genes (5,730 out of 24,482) that were selected based on their associated variance. Our procedure run on the full dataset led to the selection of 5,932 genes that fall into 22 clusters (in contrast to 154 clusters using geneMCL). This discrepancy is likely to be due to the filtering step applied to the dataset. Indeed, a strong associated variance can also be reminiscent of punctual random artifacts. Thus, selecting those genes will generate small or singleton clusters. In this context, the MDNN statistic better handle these artifacts as its purpose is to conserve genes that belong to dense region in the hyperspace.

To date, TBrowser provides user with only one partitioning solution for a dataset. However as density is heterogeneous inside a dataset, several partitioning solutions exist. For instance, if one observes a cluster containing cells of the immune system this will also frequently contain several sub-clusters that will be reminiscent of cell types (B- or T-cells for example) or activation status. Increasing MCL granularity ("Inflation" parameter) will most generally split the parent clusters and provided user with another partitioning result. However, both results can be considered as optimal and we should consider all of them. To this end we plan to propose multiple partitioning solutions for each dataset to provide a more exhaustive view of underlying biological pathways. Although, such an approach could appear computer-intensive it should be practicable, taking into account that DBF-MCL is much faster than hierarchical clustering or MCL run on a whole dataset. In addition, although we routinely obtained very relevant results with DBF-MCL, we expect that even more accurate methods will be proposed in the future.

The present work focus on human, mouse and rat Affymetrix microarray data but TBrowser can handle any type of microarrays and organism. The current release of the database already contains data obtained using other commercial (*e.g.* Agilent, Illumina Inc., GE Healthcare, Applied Biosystems, Panomics, CapitalBio Corporation, TeleChem ArrayIt, Mergen-LTD, Eppendorf Array Technologies) and non commercial platforms (*e.g.* National Cancer Institute, Vanderbilt Microarray Shared Resource, Genome Institute of Singapore), several of them being related to the MicroArray Quality Control (MAQC) project (GSE5350) [26]. However, to date, systematic analysis of all experiments performed on these platforms has not been done. The flexibility of our approach also makes it possible to integrate and compare data obtained through any kind of large scale analysis technologies providing that the experiment can be represented by a single numerical matrix (ChIP-on-chip, Protein array, large scale Real-time PCR, ChIP-seq, etc.). Three plugins (Heatmap, TBCommonGenes and TBMap) have been presented in this article but seven new plugins have been recently developed (manuscript in preparation). In the near future, the ease of plugin development will makes it possible to look for TS enriched in genes sharing transcription factor and miRNA specific motifs in their non-coding regions.

As raw data are only available for some of the microarray datasets, we used the "normalized" data provided by submitters. These data were subsequently rank-transformed and used for classification. This procedure allowed us to re-analyze a very large number of datasets. However, the drawback is that quality status of individual samples or experiments could not be determined (computing the so-called "3′/5′ ratio" requires raw data). We plan to provide extensive quality control informations through a dedicated plugin. However, we think that scientists should comply better with the MIAME guidelines and that they should provide systematically raw data when submitting a new experiment. Finally, we would like to acknowledge the GEO database team whose efforts in providing high quality repository service made this work possible.

Materials and Methods

Microarray data retrieval

Human mouse and rat microarray data derived from 30 Affymetrix microarray platforms (Supplementary Table S1) were downloaded from the GEO ftp site and retrieved in seriesMatrix file format (ftp://ftp.ncbi.nih.gov/pub/geo/DATA/SeriesMatrix/). SeriesMatrix are summary text files related to a GEO series Experiment (GSE) that include sample and experiment metadata together with a tab-delimited matrix that corresponds to normalized expression data. Each file (n = 2,869) was parsed using a Perl script to extract gene expression matrix and metadata. Probes with missing expression values were excluded from analysis. Only expression matrix with at least ten columns/samples were kept for subsequent analysis (n = 1,484, Supplementary Table S2).

DBF-MCL algorithm

The filtering step of DBF-MCL was implemented in C. The latest Markov Clustering algorithm version (1.006, 06-058) was

obtained from http://micans.org/mcl/src/. The full pipeline of DBF-MCL (that integrates normalization, filtering and partitioning) was implemented in Bash Shell Scripting language. This script supports different metrics for distance calculation (Euclidean distance, Pearson's correlation coefficient-based distance, Spearman's rank correlation-based distance).

Data normalization and processing

Given the huge amount of data processed by GEO curators it is impractical to determine the quality and efficiency of the normalization methods used [27]. Although seriesMatrix files should ideally contain log-transformed data, expression matrices in linear scale were also observed in several cases. To circumvent this problem each column of the expression matrix was rank-transform (using R software). This normalization procedure is insensitive to data distribution and provided us with a standard input for the DBF-MCL algorithm. In the case of microarray data, DBF-MCL was run using Spearman's rank correlation-based distance (1-r). However, although rank-based methods are well suited for normalization and distance calculation purposes they are not appropriate to display gene expression profiles. To this end, a normal score transformation was applied to each column of the datasets after DBF-MCL classification. The transformation ensures that whatever the data a standard format is available for heatmap visualization. Finally, for each experiment, this dataset was used (1) to classify samples using hierarchical clustering (2) to build the expression matrix for the corresponding TS.

Data storage

Expression matrix for each TS were stored in an indexed flat file with a TS ID as a key. This flat file is used by the TBrowser client to retrieve expression data for the requested TS. Experiment metadata, corresponding to sample and experiment informations were stored in a mySQL relational database. Probe meta-information (gene symbol, gene name, GenBank accession ID, chromosomal location, Entrez ID) were obtained from Bioconductor [28] annotation packages and stored in the database. In some cases, as no annotation packages were available (especially for GeneChip® CustomExpress® Array) a script was used to obtain gene symbols and gene names from GenBank files based on the provided GenBank accession ID. Both flat file and database information will be periodically updated to give access to novel experiments stored in GEO repository.

Complex9 dataset

The complexe9 dataset was obtained from the UH Data Mining and Machine Learning Group (UH-DMML, http://www2.cs.uh.edu/~ml_kdd/). Cluster Affinity Search Technique (CAST) was run using the TMEV software. QT_CLUST and k-means were run using the flexclust and fpc R package. For k-means, the algorithm was run 10 times with random initial centers. Hierarchical clustering was performed using the amap library from the R/Bioconductor project. The Euclidean distance was used in all cases.

Functional enrichment analysis

We used the DAVID knowledgebase [12] for functional enrichment analysis as it provided a practical mean to gain access to a wide range of heterogeneous sources of gene annotation (152,543 annotation terms were used for human, 105,207 for mouse and 39,787 for rat). DAVID ID mapping was obtained for 218,727 AffyID. A Perl script that integrates call to the R software was run to load probe list and calculate iteratively Fisher's exact test p-values on 2×2 contingency tables. Bonferroni adjusted *p*-values were calculated using the multtest Bioconductor library for all TS. Overall, 5.10^6 Fisher's exact test were performed.

User interface

TBrowser is accessible through a web browser at TAGC web site (http://tagc.univ-mrs.fr/tbrowser/). Of note, the TBrowser client is extensible through a plug-in architecture that allows rapid development of additional features. A developer's guide will be available soon on our website.

Supporting Information

Figure S1 A schematic overview of the pipeline used in TBrowser.
Found at: doi:10.1371/journal.pone.0004001.s001 (10.16 MB TIF)

Figure S2 An illustration in two dimensions of the motivation behind DBF-MCL filtering step. Arrows point out the 20th nearest neighbor for selected points. Length of each segment corresponds to a given DKNN value.
Found at: doi:10.1371/journal.pone.0004001.s002 (8.22 MB TIF)

Figure S3 Distributions of DKNN values. Observed DKNN values (solid line) and of a set of simulated DKNN values S (dotted line) are shown for (A) the Complex9RN200 artificial dataset and (B) the GSE1456 microarray dataset.
Found at: doi:10.1371/journal.pone.0004001.s003 (9.01 MB TIF)

Figure S4 Colors correspond to the clusters found using the corresponding algorithm (A) The whole dataset (9,112 points). (B) A zoom-in of Complex9RN200 dataset that displays the various shapes to be found. (C) DBF filtering step without partitioning. With k set to 60, noisy elements remain around the shapes. (D–G) The filtering and partitioning results obtained using DBF-MCL run with a range of k values and I values. Other arguments are unchanged (FDR = 10%, S1..3). The set of points (n = 3,108) obtained using DBF-MCL (k = 20) was used to test the other algorithms (H) Results obtained with hierarchical clustering (single linkage). The obtained dendrogram was cut to produce 9 clusters. (I) Results obtained with the QT_CLUST algorithm (radius = 0.8). (J) Results obtained for k-means (9 centers, 100 initializations). (K) Results obtained with cst(threshold = 0.81).
Found at: doi:10.1371/journal.pone.0004001.s004 (9.41 MB TIF)

Figure S5 Impact of various k values on DBF-MCL results. The x-axis correspond to k values. The y-axis correspond to the number of elements considered as informative. (A) DBF-MCL was run with the Complex9RN200 as input using a range of k values (FDR = 10%, S1..3, Inflation = 1.2). (B) DBF-MCL was run with several microarray datasets as input (including GSE1456) using a range of k values (FDR = 10%, S1..3, Inflation = 2).
Found at: doi:10.1371/journal.pone.0004001.s005 (8.72 MB TIF)

Figure S6 The TBMap plugin. These pictures are derived from the GPL96 map (22,215 probes as rows and 3,114 GPL96 specific TS as columns). Red indicates the presence of a gene in the corresponding TS (default to black). Only small parts of the map are displayed. (A) A cluster enriched in genes from the "Aminoacyl-tRNA biosynthesis" KEGG pathway (hsa00970). Genes (rows) from this KEGG pathway are displayed as blue lines (CARS, SARS, AARS, GARS, MARS, IARS, YARS). Genes from a manually entered gene list are shown in yellow (TRIB3, MOCOS, MPZL1, CBS, PPCDC). (B) A cluster enriched in genes related to oxydative phosphorylation (KEGG pathway hsa00190, "Oxidative phosphorylation"). (C) A cluster containing

genes related to ribosome biogenesis (KEGG pathway hsa03010 "Ribosome"). (D) A cluster enriched in genes involved in cell proliferation (KEGG pathway hsa04110 "Cell cycle").
Found at: doi:10.1371/journal.pone.0004001.s006 (9.66 MB TIF)

Table S1 Informations related to Affymetrix platforms (n = 33) used in the present work.
Found at: doi:10.1371/journal.pone.0004001.s007 (0.12 MB XLS)

Table S2 Informations related to experiments (n = 1,484) that were analyzed using the DBF-MCL algorithm. All Informations were obtained from the GEO website.
Found at: doi:10.1371/journal.pone.0004001.s008 (1.38 MB XLS)

Table S3 This matrix summarizes the results obtained using the "ESR1 & GATA3 & FOXA1" query. Rows correspond to genes and columns to TS. The presence of a given gene in a given TS is indicated by 1 (default 0).
Found at: doi:10.1371/journal.pone.0004001.s009 (0.66 MB XLS)

Acknowledgments

The authors would like to thanks the staff from the TAGC laboratory for helpful discussions and gratefully acknowledge Francois-Xavier Theodule for technical assistance.

Author Contributions

Conceived and designed the experiments: FL JT AB GD ER SG DP. Performed the experiments: FL JT AB DP. Analyzed the data: FL JT AB JI CN DP. Contributed reagents/materials/analysis tools: FL JT AB GD ER SG JI DP. Wrote the paper: JI CN DP.

References

1. Stoeckert CJ, Causton HC, Ball CA (2002) Microarray databases: standards and ontologies. Nat Genet 32 Suppl: 469–73.
2. Barrett T, Edgar R (2006) Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol 411: 352–69.
3. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, et al. (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Res 31: 219–23.
4. Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, et al. (2005) ArrayExpress–a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 33: D553–5.
5. D'haeseleer P (2005) How does gene expression clustering work? Nat Biotechnol 23: 1499–501.
6. Heyer LJ, Kruglyak S, Yooseph S (1999) Exploring expression data: identification and analysis of coexpressed genes. Genome Res 9: 1106–15.
7. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling (n.d.) Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi = 10.1.1.44.5847. Accessed 18 September 2008.
8. Van Dongen S (2000) A cluster algorithm for graphs. National Research Institute for Mathematics and Computer Science in the. pp 1386–3681.
9. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637–43.
10. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575–84.
11. Samuel Lattimore B, van Dongen S, Crabbe MJC (2005) GeneMCL in microarray analysis. Comput Biol Chem 29: 354–9.
12. Sherman BT, Huang DW, Tan Q, Guo Y, Bour S, et al. (2007) DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics 8: 426.
13. Pawitan Y, Bjöhle J, Amler L, Borg A, Egyhazi S, et al. (2005) Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res 7: R953–64.
14. Lacroix M, Leclercq G (2004) About GATA3, HNF3A, and XBP1, three genes co-expressed with the oestrogen receptor-alpha gene (ESR1) in breast cancer. Mol Cell Endocrinol 219: 1–7.
15. Rogers MA, Langbein L, Winter H, Ehmann C, Praetzel S, et al. (2001) Characterization of a cluster of human high/ultrahigh sulfur keratin-associated protein genes embedded in the type I keratin gene domain on chromosome 17q12-21. J Biol Chem 276: 19440–51.
16. Johansson P, Pavey S, Hayward N (2007) Confirmation of a BRAF mutation-associated gene expression signature in melanoma. Pigment Cell Res 20: 216–21.
17. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863–8.
18. Strubin M, Newell JW, Matthias P (1995) OBF-1, a novel B cell-specific coactivator that stimulates immunoglobulin promoter activity through association with octamer-binding proteins. Cell 80: 497–506.
19. Zhao C, Inoue J, Imoto I, Otsuki T, Iida S, et al. (2008) POU2AF1, an amplification target at 11q23, promotes growth of multiple myeloma cells by directly regulating expression of a B-cell maturation factor, TNFRSF17. Oncogene 27: 63–75.
20. Rabot M, El Costa H, Polgar B, Marie-Cardine A, Aguerre-Girr M, et al. (2007) CD160-activating NK cell effector functions depend on the phosphatidylinositol 3-kinase recruitment. Int Immunol 19: 401–9.
21. Boles KS, Nakajima H, Colonna M, Chuang SS, Stepp SE, et al. (1999) Molecular characterization of a novel human natural killer cell receptor homologous to mouse 2B4. Tissue Antigens 54: 27–34.
22. PubMed Snapshot (n.d.) Available: http://www.ncbi.nlm.nih.gov/sites/entrez. Accessed 18 September 2008.
23. Townsend MJ, Weinmann AS, Matsuda JL, Salomon R, Farnham PJ, et al. (2004) T-bet regulates the terminal maturation and homeostasis of NK and Valpha14i NKT cells. Immunity 20: 477–94.
24. Lauwerys BR, Renauld JC, Houssiau FA (1999) Synergistic proliferation and activation of natural killer cells by interleukin 12 and interleukin 18. Cytokine 11: 822–30.
25. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–6.
26. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, et al. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 24: 1151–61.
27. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. Nucleic Acids Res 35: D760–5.
28. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.

4.5 Mise à jour de la base et intégration de données

Au début de ma thèse, une mise à jour de la base de données a été nécessaire afin d'améliorer les performances et la cohérence des données pour inclure plus d'expériences, d'espèces et d'annotations. En effet, les données présentes dans la base de données avaient été récupérées en 2007 au début du projet. Or la quantité d'échantillons disponibles a pratiquement doublé en 2 ans (Figure 4.3).

Enfin, une fois la preuve de concept réalisée, nous avons voulu conforter nos observations en incluant d'autres sources de données, afin d'aider l'utilisateur à construire des réseaux de gènes contextualisés en fonction d'une pathologie, d'une voie de signalisation ou bien encore d'un tissu donné.

4.5.1 Restructuration de la base de données

Devant la nécessité d'améliorer les performances de TBrowser, tout particulièrement au niveau de la rapidité d'exécution des requêtes sur la base de données, le schéma de cette base a été totalement redéfini. En effet, la première version de la base de données contenait des informations redondantes et près de 200 tables non indexées (avec notamment une table par plateforme). J'ai donc dû normaliser les tables existantes afin de diminuer la redondance et permettre un accès plus rapide aux données.

Pour réduire cette redondance et permettre une meilleure gestion des données, j'ai mis au point une base de données MySQL (version 5.0) avec un moteur de stockage de type MyISAM. Ce moteur ne supporte pas les transactions (regroupement de plusieurs instructions en une seule), ni les clefs étrangères (contrainte d'intégrité de la base de données). Cependant, j'ai choisi ce type de moteur car celui-ci est simple à mettre en oeuvre et car il est généralement conseillé pour les applications utilisant essentiellement des requêtes en lecture et donc peu de requêtes en écriture. L'absence de transactions a été remplacée par l'utilisation de fonctions sous forme de procédures stockées. La gestion de l'intégrité des tables a été implémentée au niveau de l'application TBrowser. Cependant afin de limiter la redondance des données, des clefs étrangères ont été créées telles que « expID » pour qualifier une expérience et « signatureID » pour une signature transcriptionnelle. Les identifiants de gènes ont tous été rapprochés de l'identifiant le plus utilisé et surtout disponible pour toutes les espèces sous le même format de plus petite taille, les gene ID (valeur numérique, facile à stocker dans la base et noté entrezID dans notre base) (Figure 4.4).

L'utilisation des fichiers plats indexés contenant les matrices d'expression a été conservée, car il n'existe pas à ma connaissance de moyen plus efficace de stocker ce type de données. Leur inclusion dans la base de données aurait entraîné un alourdissement de celle-ci et donc des performances médiocres. Cependant toutes les signatures ont été re-extraites en utilisant la librairie R RTools4TB, que j'ai développée et qui sera présentée en détail plus loin.

TBrowser a été modifié en conséquence afin de prendre en compte la nouvelle base de don-

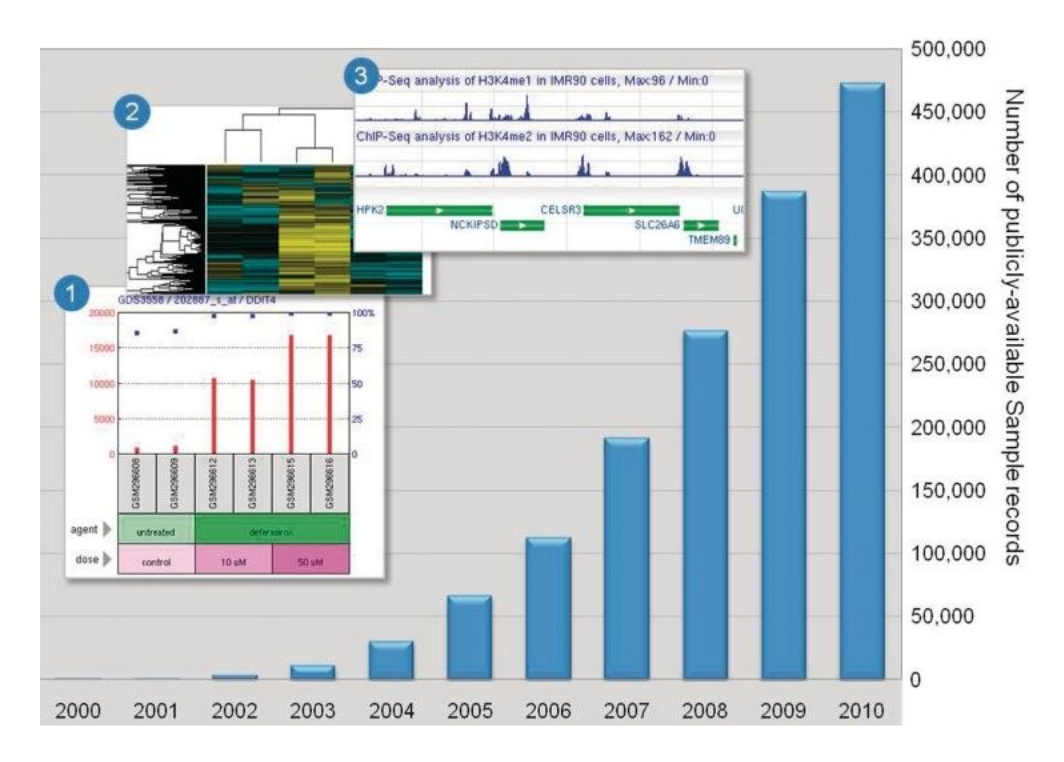


FIGURE 4.3 – Évolution du nombre d'échantillons disponibles dans Gene Expression Omnibus de 2000 à 2010. (Adapté de [Barrett *et al.*, 2005])

nées, mais également l'inclusion dans cette base d'une centaine de procédures stockées non présentes dans la première version et qui nous permettent de simplifier la récupération des résultats de requête au niveau de l'application java. Un seul appel à une procédure stockée est maintenant suffisant, là où plusieurs requêtes SQL étaient nécessaires.
De plus, l'utilisation d'une base de données installée localement, contenant des expériences non publiées dans GEO, a été réalisée pour permettre à un utilisateur avancé, de pouvoir créer sa propre base de données. La base de données a été installée au CRG (Center for Genomic Regulation) de Barcelone et au sein de l'IGF (Institut de Génomique Fonctionnelle) de Montpellier.

4.5.2 Intégration de nouvelles données

Afin de pouvoir inclure beaucoup plus d'espèces, d'autres plateformes ont été récupérées sur le serveur FTP de GEO. Contrairement à la première version de la base de données, les informations disponibles ne sont pas centrées sur les sondes mais sur les gènes par le biais des gene ID, valeur numérique et unique pour chaque gène.

L'utilisation des gene ID comme référence de gène, en suivant l'exemple du NCBI, nous a permis d'inclure d'autres types d'identifiants permettant l'accès simplifié à d'autres bases de données telles que : Uniprot, Ensembl, UCSC, Refseq, OMIM. Une table contenant les alias d'un gène a également été créée afin de permettre à l'utilisateur de rentrer une liste de gènes ne contenant pas forcément uniquement les identifiants de gènes officiels (HUGO pour Homo sapiens). Enfin, ceci pourra à terme nous permettre de mettre à jour les informations sur les gènes à partir des gene ID (c'est-à-dire les geneSymbol et autres alias du gène) sans modifier la composition des signatures.

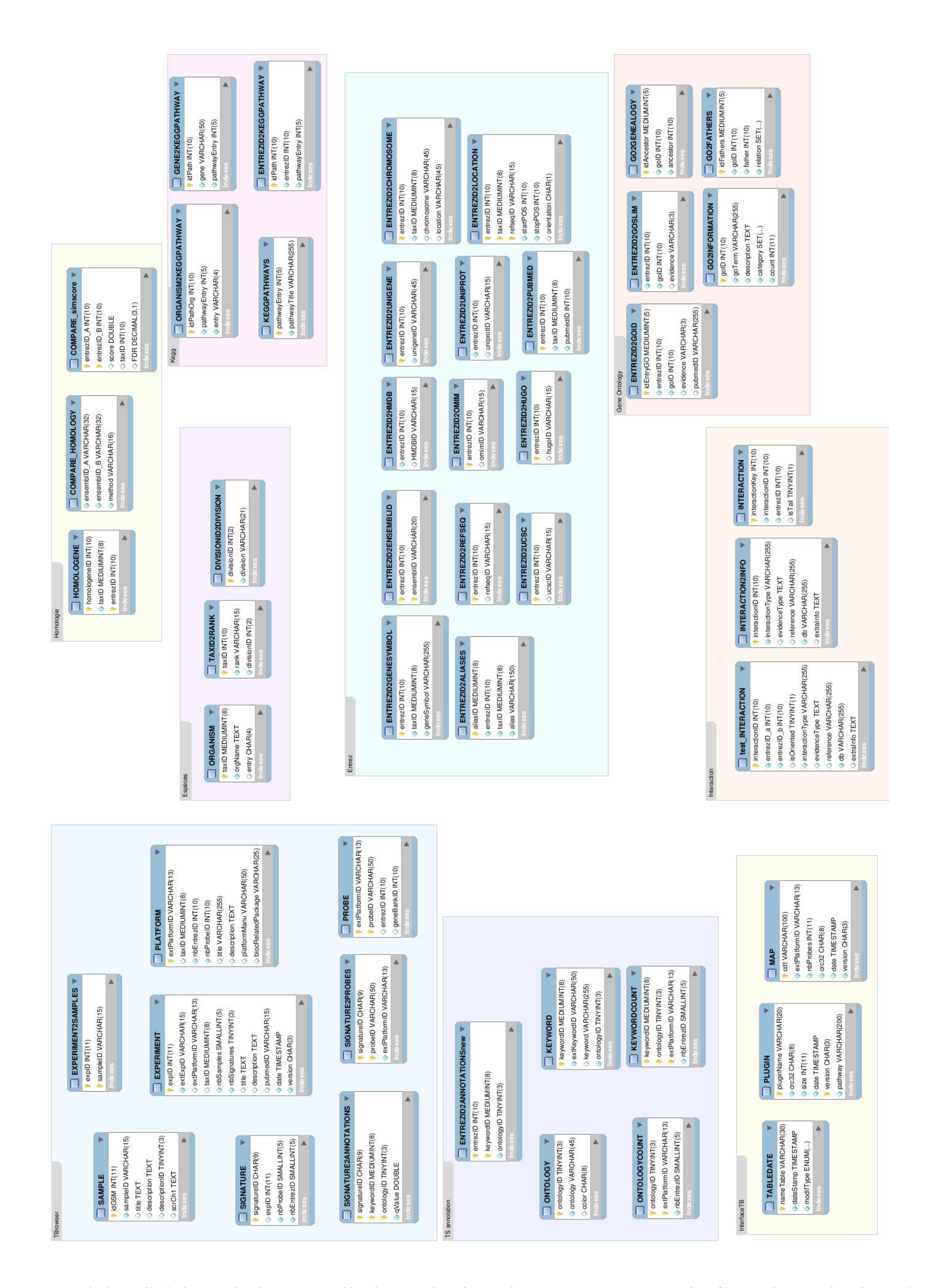


FIGURE 4.4 – Schéma de la nouvelle base de données TBrowserDBv2. Cette base de données MySQL est constituée de 47 tables d'architecture MyISAM, permettant un accès rapide aux données

miRNA
PICTAR_4WAYS
PICTAR_5WAYS
PICTAR_CHICKEN
PICTAR_DOG
TARGETSCAN_HS
TARGETSCAN_MM
TARGETSCAN_WORM
TARGETSCAN_DROSO

MOTIFS
TFBS_CONSERVED
ECRbase
CISRED

PATHWAYS
KEGG_REACTION
NCICB_CAPATHWAY
KEGG_COMPOUND
PANTHER_FAMILY
REACTOME
BBID
PANTHER_PATHWAY
BIOCARTA
PANTHER_SUBFAMILY
PANTHER_TERM_BP
KEGG_PATHWAY
PANTHER_TERM_MF
WIKIPATHWAY

GENOMIC LOCATION
CHROMOSOME
CYTOBAND

MICROARRAY DB
GENESIGDB
MSIGDB
TB_GENESETS

GENE ONTOLOGY
GOTERM_MF_ALL
GOTERM_CC_ALL
GOTERM_BP_ALL

LITERATURE
PUBMED_ID
HIV_INTERACTION_PUBMED_ID

HIV
HIV_INTERACTION
HIV_INTERACTION_CATEGORY

DISEASE
OMIM_PHENOTYPE
OMIM_ID
DISEASE
PHENOTYPE
GENETIC_ASSOCIATION_DB

PROTEIN DOMAINS
SMART_NAME
COG_KOG_NAME
TIGRFAMS_NAME
PRODOM_NAME
SCOP_ID
PROSITE_NAME
COG_KOG_ONTOLOGY
PIR_SUPERFAMILY_NAME
PFAM_NAME
KEA
SP_PIR_KEYWORDS
HMDB

FIGURE 4.5 – Les diverses annotations disponibles dans la base de données utilisée pour générer l'annotation des signatures transcriptionnelles

De nombreuses bases de données d'annotation utilisent également les gene ID comme identifiant de gène, ceci nous a permis d'inclure d'autres sources d'annotation afin de pouvoir analyser les signatures transcriptionnelles (Figure 4.5). De plus, par recoupement avec les nouveaux identifiants contenus dans la base de données, nous avons également pu intégrer des annotations dont les identifiants étaient des noms de gènes (parfois non officiels), des EnsemblID, des UniprotID. J'ai ainsi créé des procédures stockées et un script bash permettant (1) d'extraire les données, (2) de les formater pour leur intégration dans la base de données et (3) de remplir toutes les tables concernant les annotations (KEYWORD, KEYWORDCOUNT, ONTOLOGY, ONTOLOGYCOUNT).

J'ai ensuite implémenté un autre script bash, faisant appel à la base de données et à R, pour permettre le calcul des enrichissements fonctionnels pour une annotation et une liste de signatures.

Finalement, pour pouvoir facilement et automatiquement générer les signatures transcriptionnelles, un pipeline automatique a été réalisé à partir des divers programmes en R, perl, C et bash développés au laboratoire.

Il permet à partir d'une liste de GSE d'extraire les matrices d'expression normalisées, de vérifier maintenant si on dispose de la GPL dans notre base de données, de filtrer les expériences de plus de 8 échantillons, et bien sûr de générer automatiquement les signatures (à l'aide de notre librairie R/Bioconductor RTools4TB). Une fois toutes ces informations récoltées, les données sont automatiquement incluses dans la base de données.

Ce pipeline permet de pouvoir inclure n'importe quelle expérience, à condition que la plateforme soit présente dans notre base de données et que l'on dispose du fichier « seriesMatrix » provenant de GEO. De plus, l'utilisation de procédures stockées nous permet également de remplir toutes les tables automatiquement. L'enrichissement fonctionnel des signatures transcriptionnelles est ensuite validé par un test exact de Fisher, avec une correction de Benjamini et Hochberg.

Toutes ces nouvelles données sont accessibles via l'interface graphique de TBrowser mais

également par de nouveaux plugins, des services web et une librairie R/Bioconductor, ainsi que par le développement de nouveaux plugins présentés ci-après.

4.6 Développement de nouvelles fonctionnalités

4.6.1 Nouveaux modes de requêtes

Afin de répondre aux demandes des utilisateurs, il est maintenant possible de construire des requêtes booléennes basées sur les identifiants gene ID et homologene ID. De plus, partant du principe que l'on ne retrouve pas forcément à chaque fois l'ensemble des gènes comme étant coexprimés, nous avons implémenté un autre mode de requête, non booléen, par liste. Cette variabilité des signatures peut également être visualisée grâce aux cartes transcriptionnelles dans le plugin TBMap.

Ce nouveau mode de requête permet à l'utilisateur d'entrer une liste de gènes par exemple, par un simple copier-coller dans la zone de requête, et de demander quelles sont les signatures comportant au minimum une certaine proportion des gènes de cette liste. Il peut ainsi être utilisé à partir d'un groupe de gènes trouvés comme différentiellement exprimés dans une étude de puce à ADN, ou bien à partir de gènes cibles d'un facteur de transcription donné provenant d'une étude épigénétique par ChIP-seq.

4.6.2 Amélioration et nouveaux plugins

Depuis la publication de TBrowser en 2008, quelques améliorations ont été apportées aux plugins existants :

- Heatmap (développé par Fabrice Lopez)(Figure 4.6) ;
- TBNeighborhood (anciennement TBCommonGenes) avec l'ajout des informations sur les gènes (différents identifiants et annotations) qui n'étaient pas accessibles dans la première version de la base de données (Figure 4.6) ;
- TBMap (développé par Fabrice Lopez), j'ai modifié le plugin pour permettre l'inclusion de cartes transcriptionnelles provenant de nouvelles espèces (autre que Homme et souris) ; d'autres améliorations ont également été réalisées : visualisation des gènes appartenant à une voie de signalisation KEGG, zoom à partir de sélection sur la carte transcriptionnelle, corrélation gène-gène ; Plusieurs nouveaux plugins ont été développés (Figure 4.6) :
- AnnotationOverview permettant de visualiser les annotations communes d'une liste de signatures ;
- KeggSearch (développé par Fabrice Lopez) représentant les pathways KEGG associés à une liste de gènes provenant par exemple de TBNeighborhood. J'ai réalisé une procédure stockée permettant la génération du tableau de résultat affiché dans le plugin plutôt que de traiter ces données au niveau du plugin java, ceci en incluant dans notre base de données les informations sur tous les pathways Kegg (à partir de leur base de données) (Figure 4.6) ;

- TBConvertor permettant l'obtention de tous les identifiants et informations d'une liste de gènes à partir d'une liste d'identifiants (gene ID, GeneSymbol, homologene ID) ;
- TBMotifsSearch pour interroger les outils de recherche de motifs cis-régulateurs, TFM-Explorer et DiRE, à partir d'une signature ou d'une liste de gènes ;
- InteractomeBrowser (développé par Cyrille Lepoivre) représente une liste de gènes sous forme d'interactome protéine-protéine, en y ajoutant des informations sur les cibles de miARN et sur les gènes cibles de facteurs de transcription. J'ai ainsi intégré dans la base de données ces sources d'annotation (Figure 4.6).

Ce dernier plugin a fait l'objet d'une publication acceptée dans BMC Bioinformatics et a été utilisé pour représenter des résultats d'une analyse transcriptome d'infection par *Coxiella burnetii* de souris dans une publication par Textoris et collaborateurs en août 2010 ([Textoris *et al.*, 2010]).

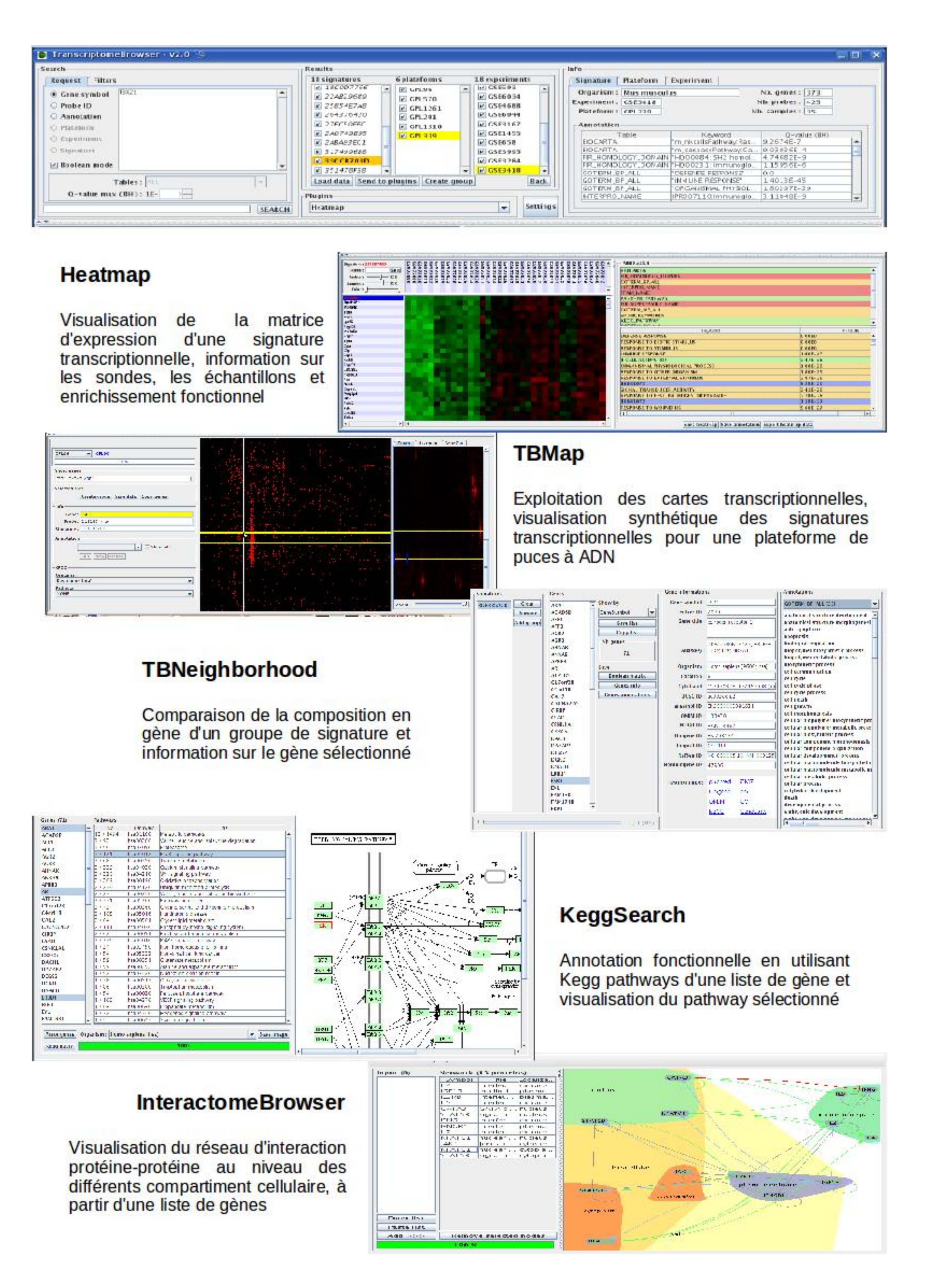


FIGURE 4.6 – Interface graphique de TBrowser avec son panneau de requêtes et ses principaux plugins

TranscriptomeBrowser 3.0: introducing a new compendium of molecular interactions and a new visualization tool for the study of gene regulatory networks

Cyrille Lepoivre[1,2], Aurélie Bergon[1,2], Fabrice Lopez[1,2,3], Narayanan B Perumal[4], Catherine Nguyen[1,2,3], Jean Imbert[1,2,3] and Denis Puthier[1,5§]

[1] Inserm UMR_S 928, TAGC, Parc Scientifique de Luminy, Marseille, France

[2] Université de la Méditerranée, Marseille, France

[3] IBiSA Platform TGML, Parc Scientifique de Luminy, Marseille, France

[4] Eli Lilly and Company, Indianapolis, Indiana, USA

[5] ESIL, Universités de Provence et de la Méditerranée, Marseille, France

[§]Corresponding author

Email addresses:

CL: lepoivre@tagc.univ-mrs.fr

AB: bergon@tagc.univ-mrs.fr

FL: lopez@tagc.univ-mrs.fr

NBP: nperumal@iupui.edu

CN: nguyen@tagc.univ-mrs.fr

JI: jean.imbert@inserm.fr

DP: puthier@tagc.univ-mrs.fr

Abstract

Background:

Deciphering gene regulatory networks by *in silico* approaches is a crucial step in the study of the molecular perturbations that occur in diseases. The development of regulatory maps is a tedious process requiring the comprehensive integration of various evidences scattered over biological databases. Thus, the research community would greatly benefit from having a unified database storing known and predicted molecular interactions. Furthermore, given the intrinsic complexity of the data, the development of new tools offering integrated and meaningful visualizations of molecular interactions is necessary to help users drawing new hypotheses without being overwhelmed by the density of the subsequent graph.

Results:

We extend the previously developed TranscriptomeBrowser database with a set of tables containing 1,594,978 human and mouse molecular interactions. The database includes: (i) predicted regulatory interactions (computed by scanning vertebrate alignments with a set of 1,213 position weight matrices), (ii) potential regulatory interactions inferred from systematic analysis of ChIP-seq experiments, (iii) regulatory interactions curated from the literature, (iv) predicted post-transcriptional regulation by micro-RNA, (v) protein kinase-substrate interactions and (vi) physical protein-protein interactions. In order to easily retrieve and efficiently analyze these interactions, we developed InteractomeBrowser, a graph-based knowledge browser that comes as a plug-in for TranscriptomeBrowser. The first objective of InteractomeBrowser is to provide a user-friendly tool to get new insight into any gene list by providing a context-specific display of putative regulatory and physical interactions. To achieve this, InteractomeBrowser relies on a "cell compartments-based layout" that makes use of a subset of the Gene Ontology to map gene products onto relevant cell compartments. This layout is particularly powerful for visual integration of heterogeneous biological information and is a productive avenue in generating new hypotheses. The second objective of Interactome-

Browser is to fill the gap between interaction databases and dynamic modeling. It is thus compatible with the network analysis software Cytoscape and with the Gene Interaction Network simulation software (GINsim). We provide examples underlying the benefits of this visualization tool for large gene set analysis related to thymocyte differentiation.

Conclusions

The InteractomeBrowser plugin is a powerful tool to get quick access to a knowledge database that includes both predicted and validated molecular interactions. InteractomeBrowser is available through the TranscriptomeBrowser framework and can be found at : http://tagc.univ-mrs.fr/tbrowser/. Our database is updated on a regular basis.

Introduction

In the last decade, the advent of high throughput technologies led to the emergence of the systems biology era and prompted the research community to systematically define the expression levels of mRNAs and micro-RNA (miRNAs) through thousands of cell and tissues under physiological and pathological conditions [1]. Now, one of the crucial issues is to define the biological mechanisms that drives genes expression with the ultimate goal of reverse-engineering gene regulatory networks (GRN) as a whole in order to predict the system outcome under molecular perturbations.

One current limit for biologists interested in mining regulatory information or for bioinformaticians interested in creating regulatory maps for modeling, is that this information is scattered over the Internet under various formats making it difficult to handle. Thus one needs to create a unified database that would list known and predicted molecular interactions. This information can be obtained from different sources: (i) from the literature, (ii) from large-scale experimental methods that allow genome-wide profiling of transcription factors (TFs) binding sites to DNA or (iii) from DNA sequence analysis, by searching 3'UTR regions for miRNA specific motifs or by scanning gene promoters with transcription factor specific position weight matrices (PWMs). In the latter case, the use of comparative genomics is known to greatly improve predictions of functional TF binding sites by limiting the number of false positives (though increasing false negative rate) [2,3]. Another limit of GRN analysis is the intrinsic complexity of the data. In this regard, several graph-based tools have been developed to draw a global picture of the putative interactions taking place in the biological context of interest (for a review, see reference [4]). In these, genes or proteins appear as nodes in a graph, and functional relations (physical/regulatory interactions) are represented as edges connecting the corresponding entities. The topology of the subsequent network can later be analyzed using advanced tools such as Cytoscape [5]. However, as data integration is a challenge that requires to map various types of evidence onto a set of stable gene ids, most applications are oriented toward a single data type (mostly regulatory or physical interactions, see table 1 for an overview) [6-10]

Moreover, another challenge is the development of graph-based tools producing clear, meaningful and integrated visualizations from which users can draw new hypotheses without being overwhelmed by the density of the presented graphic information. In this regard, the Cytoscape plug-in "Cerebral" proposes an intuitive visualization method through a "cell compartment-based layout" that shows interacting proteins on a layout resembling "traditional" signalling pathway/system diagrams [11].

Here, we sought to create a compendium of predicted and validated molecular interactions in human and mouse. First, we used a large collection of PWMs obtained from TRANSFAC (n=523) , JASPAR (n= 303) and UNIPROBE (n=387) to search, in gene promoter regions, for candidate transcription factor binding sites (TFBSs) conserved over human, mouse, rat and dog genomes [12-14]. Overall, our analysis of these PWMs corresponding to 347 human and 475 mouse transcription factors (TFs) provides a systematic overview of gene regulation in the human and mouse. Data generated in this study were next integrated with a large set of molecular interactions from various sources including (i) potential protein/DNA interactions derived from ChIP-seq experiments (ChIP-X database), (ii) curated regulatory interactions obtained from the literature (OregAnno, LymphTF-DB), (iii) predicted miRNA/targets interactions (TargetScan) (iv) protein kinase-substrate interactions derived from multiple online sources (KEA) and (v) physical protein-protein interactions obtained from IntAct and HPRD [15-21]. Informations related to these interactions were stored as MySQL tables that were integrated in the back-end database of TranscriptomeBrowser, our previously published microarray datamining software [22]. Finally, we developed InteractomeBrowser (IBrowser) as a plugin for TranscriptomeBrowser. IBrowser was developed using the prefuse Java library and can be used to translate any gene list into a meaningful graph. The specificity of the IBrowser plugin relies on a new "cell compartments-based layout" that makes use of a subset of the Gene Ontology to map gene products onto relevant cell compartments. This layout is particularly powerful for visual integration of heterogeneous biological information. Moreover, IBrowser is in-

tegrated into the TranscriptomeBrowser suite, which allows an easy communication with other tools, for instance to retrieve lists of genes that are frequently coexpressed in given conditions, thus creating context-specific views of the interactome and regulome.

IBrowser is intended both for biologists and bioinformaticians. On one hand, it is a graph-based knowledge browser, that is intended to provide new insight into any user-defined gene list. On the other hand it is also intended to fill the gap between heterogeneous genomic data and gene regulatory network analysis. In this regard, graphs produced inside IBrowser may be exported into Cytoscape and GINsim, a dynamic modeling software [23]. In the following sections we provide several examples underlying the benefits of this visualization tool for large gene set analysis.

Implementation

We first used phylogenetic footprinting to predict regulatory elements in the human and mouse genomes. A dataset of 1,213 PWMs corresponding to mouse or human transcription factors was obtained from various sources (TRANSFAC 10.2, JASPAR 2010, UNIPROBE). The multiz28way (with hg18 as a reference) and the multiz30way (with mm9 as a reference) cross-species multiple alignments were obtained from UCSC [24]. We retained for analysis alignments flanking transcription start sites on both sides (-3000, 3000) of any RefSeq transcript and devoid of coding sequences. Sequences were scored following the commonly used formula [25]:

$$SCORE_{p,c} = \sum_{w=0}^{W-1} log2\left(\frac{P\left(seeing\ S_{p+w}\ at\ position\ w \mid PWM\right)}{P\left(seeing\ S_{p+w}\ at\ position\ w \mid Background\ model\right)}\right)$$

where $SCORE_{p,c}$ represents the PWM score for a PWM of length W in the DNA sequence of a species c between positions p and $p+W-1$ and S_{p+w} represents the nucleotide observed at position $p+w$. The probability of observing each nucleotide under the background distribution was assumed to be 0.25. For each PWM m, a score threshold t_m with p-value below 5.10^{-5} was computed using

matrix-distrib from RSAT ensuring high stringency of sequence scoring [26]. A sequence in the reference genome was considered as a putative TFBS if its score for PWM *m* at position *p* in the alignment was found above t_m in human, mouse rat and dog. Each PWM was then linked to its corresponding transcription factors and putative targets. Information was stored in a MySQL relational database.

We also integrated several informations obtained from popular databases. Protein/DNA interactions (n= 174,168) derived from various genome wide analysis (e.g.; ChIP-on-chip, ChIP-seq and ChIP-PET) and encompassing interactions corresponding to 38 human TFs and 55 mouse TFs were obtained from the ChIP-X database. TFBS predictions were obtained from the present work (see below) and TFBSConserved UCSC track (367,829 and 686,936 respectively). A set of regulatory interactions curated from the literature were obtained from LymphTF-DB (392 directed interactions) and OregAnno (1,991 interactions). Protein-protein interaction datasets were obtained from HPRD and Intact (39,224 and 50,286 respectively). Protein kinase-substrate relationships were retrieved from KEA (n=14,084). Finally, miRNA/target relationships were obtained from TargetScan database predictions (n=260,068). For all datasets, all identifiers were mapped onto Entrez Gene ids. This compendium of molecular interactions is available as flat files at: *ftp://tagc.univ-mrs.fr/public/TranscriptomeBrowser/DB_Tables/*.

InteractomeBrowser was developed using the Prefuse Java library which was modified according to our needs. InteractomeBrowser requires Java 1.6.

Results and discussion

TFBS predictions using comparative genomics

Although previous works have demonstrated the power of comparative genomics in defining novel regulatory motifs in human and mouse, few of them integrate the PWMs recently computed from protein binding microarray (PBM) experiments. Overall, restricting our analysis to promoter regions and using a set of 1,213 PWMs, we predicted TFBSs in 141,305 position-specific motifs of the mouse genome and 164,171 of the human genome. The median number of hits for any PWM was 117 in mouse (mean, 169; range, 3-2,317) and 122 in human (mean, 192; range, 6-2,678). The PWMs with highest number of hits correspond to Sp1 transcription factor (M00931, M00933, M00196) in both species (Supplementary Figure S1). Sp1 binds GC-rich elements (consensus, GGGGCGGGGC) that are found in the promoter regions of a large number of genes [27]. As promoter regions are known to contain CpG islands we checked whether our approach could overestimate the number of targets for TF with high GC-content related PWMs. As shown in figure S1, this effect was essentially restricted to Sp1 and to a lesser extend to the Maz related PWM (consensus, RGGGAGGG). As expected, PWMs with high information content were most generally associated with fewer motifs (Figure S1, point size).

Genes with highly conserved promoter regions mostly encode transcription factors

We next estimated the number of predicted regulators for each gene by computing the number of non-redundant PWMs associated with each gene. The number of PWMs that have a significant match in gene promoter regions range from 1 to 318 (median, 8; mean, 13.37) in mouse and 1 to 353 in human (median, 7; mean 13.17). Genes in the top 1% considering the number of regulators (eg; Lmo3, Foxp2, Bcl11a) were, as expected, invariably associated with highly conserved promoter regions. Moreover, functional annotation indicates that a very large proportion of these genes were transcription factors and genes related to development. Indeed, in mouse, enrichment analysis

of the gene list (112 genes) using Fisher's exact test (with Benjamini and Hochberg correction) indicated a very strong enrichment for genes related to terms “Transcription factor” (PANTHER TERM; q-value, $1.3.10^{-27}$; 52 genes out 95 annotated), “pattern specification process” (GO biological process; q-value, $2.8.10^{-13}$; 19 genes out 78 annotated) or “neuron differentiation” (GO biological process; q-value,$1.48.10^{-09}$; 18 genes out 78 annotated). Very concordant results were also observed for human (a summary of functional enrichment analysis using the ClueGO cytoscape plugin is provided in Supplementary Figure S2 and S3) [28]. Actually, these results are in agreement with the work of Bejerano and collaborators that showed that ultraconserved elements of the human genome are most often found in genes involved in the regulation of transcription and development [29]. As a consequence our phylogenetic footprinting analysis predicts a higher number of motifs in the promoter regions of these genes. Although TFBS conservation in mammals has been previously analyzed in several papers, none of them, to our knowledge, reported this observation that may introduce a bias in the analysis. However, these ultraconserved regions may also be reminiscent of HOT (high-occupancy target) regions identified using ChIP-seq analysis in *Caenorhabditis elegans* and Drosophila [30,31]. Indeed, HOT regions have been shown to be significantly associated with “essential genes” (*i.e.*; having an RNAi phenotype of 100% larval arrest, embryonic lethality, or sterility) and genes related to growth, reproduction, and larval and embryonic development. However, we cannot rule out that these ultra-conserved regions may be also related to other mechanisms than regulation by site-specific TFs

Biological relevance of the TFBS predictions

One criterion to assess the reliability of our predictions is based on the hypothesis that the overall functional properties of the predicted targets can be used to infer the biological processes in which TFs are involved. To test this hypothesis, we used annotation terms obtained from GO (biological process), KEGG, PANTHER, PFAM, SMART, PROSITE, and WIKIPATHWAYS databases and performed systematic annotation of all predicted target sets in the mouse [32]. For each pair of

term/PWM we computed the Fisher's exact test p-value *f*. Each cell of a matrix with terms (n=3,905) as row and PWM (n=1,103) as column was filled with a score defined as *-log(f)*. We then searched for biclusters inside this matrix using “the binary inclusion maximal algorithm “ (BiMax) [33]. Given the amount of information produced by this analysis, only some meaningful results will be presented and are summarized in figure 1. Sites for PWM related to ETS (M00746, M00971, M00771, M00339, MA0136, M00658, M00678), STAT, IRF and RUNX (M00722) transcription factor families, known to contribute to pathogen responses, were significantly over-represented in genes annotated as “immune system process” and “lymphocyte activation” (Figure 1A). Sites for PWMs related to the Rel/NF-κB pathway were significantly associated with targets related to “induction of apoptosis”, “Toll-like receptor signaling pathway” and, as expected to “NF-kappaB cascade” (Figure 1B). More subtle biclusters related to immune system were also found. As an example, RBPJK specific PWMs (M01112, M01111) were statistically significantly associated with terms “Notch signaling pathway”. Although RBPJK is already known to be crucial in NOTCH signaling pathway, PWMs related to TCF3 (also known as E2A and E47) and AP-4 were also found in the same bicluster (Figure 1C). This observation is very consistent with the known role of these TFs in early B-cell differentiation, a development step for which Notch pathway is decisive [34,35]. As expected, a bicluster containing almost all E2F-related PWMs was also found. Finally, several biclusters related to “Muscle contraction”, “Phosphorus metabolic processes”, “Synaptic transmission”, “Protein catabolic processes” and “Pre-mRNA processing” were also observed and are presented in figure 2E-I. Altogether, these results highlight the biological relevance of the TFBS predictions and provides a systematic overview of putative regulatory interactions in human and mouse. These predictions have been termed “TBMC” (TranscriptomeBrowser Motif Conservation) and are available through the InteractomeBrowser plugin or as a bed file (See supplementary material).

InteractomeBrowser: graph-based knowledge browser

The InteractomeBrowser application can be used to connect to our database in order to identify and analyze molecular interactions (See supplementary material for a video tutorial). Available molecular interactions are derived from various sources: our predictions (TBMC) and numerous databases including ChIP-X, LymphTF-DB, OregAnno, HPRD, Intact, TargetScan and KEA. However, InteractomeBrowser may also accept additional interaction datasets that users can provide through a tabulated flat file.

InteractomeBrowser relies on a mixed graph that contains both directed and undirected edges, depicting various types of interactions ranging from proteins complex formation to transcriptional regulation. Thus nodes represent both genes and gene products.

InteractomeBrowser uses a subset of terms of the Cellular Component ontology (supplementaty figure S4) to map nodes onto a schematic and hierarchical view of cell compartments (users may choose to disable this option). As a consequence, each gene product may be represented by several instances (*e.g.* : one in the nucleus and one in the cytosol).

The nodes placement is controlled by a force-directed placement layout: the nodes are repulsive to each other, they are attracted to their respective compartments, and edges act like springs (the force-directed placement layout can be switched off or on at any moment through the “Display” menu). Once a graph has been drawn, one can easily add or delete nodes. InteractomeBrowser provides several filters that are intended to focus on the most interesting part of the network. Users can filter out orphan nodes and empty compartments. An option called "Hide intercompartmental edges" allows users to remove several unlikely edges of the network, notably those involving physical interactions between distant compartments (eg; an instance of gene A in the nucleus and an instance of gene B in the extracellular regions). When the mouse is over a node or an edge, corresponding information is provided in the "Infos" tab on the left side of the application. Right-clicking on a node opens a context menu, allowing users to (i) open the NCBI web page for this gene, (ii) add regulato-

ry interactions involving this gene and other genes of the network, (iii) move the node to another compartment and (iv) connect to UCSC genome browser. The action menu provides other tools to expand the network: (i) add all the interactors of the selected genes or (ii) add common interactors of selected genes.

IBrowser can be used with any user-defined gene list, for examples genes of interest in a particular experiment. Additionally, the integration of this tool into the TranscriptomeBrowser suite facilitates the analysis of lists corresponding to pre-processed clusters of co-expressed genes stored in the database.

The next part of the result and discussion section demonstrates the use of InteractomeBrowser for retrieving molecular interactions in the context of thymocyte differentiation analysis.

Case study: early T-cell development in mouse

The development of mature T cells from lymphoid progenitor cells involves a series of cell fate choices that direct differentiation. In the context of the Immunological Genome Project (ImmGen), M.W. Painter *et al* used rigorously standardized conditions to analyze expression levels of protein-coding gene in almost all defined T-cell populations of the mouse [36]. Using SAM analysis (FDR 15%), we selected a set of 281 genes repressed during the transition from thymic DN3 stage to DN4 stage. Careful analysis, indicated that this gene set was highly enriched in genes previously shown to be crucially involved during the first step of thymocyte development. This includes cell surface markers such as Il2ra/Cd25, and Il7r together with several transcriptional regulators, including Notch1, Smarca4/Brg1, Dtx1/Deltex1, and Hes1/Hry. More recently, Neilson *et al* identified specific miRNAs enriched at distinct stages of thymocyte development by deep sequencing [37]. The authors showed that transcripts of the mir17 family are up-regulated at DN4 stage and thus could be involved in the repression of DN3 specific messenger RNAs during DN3 to DN4 transi-

tion. We thus combined one member of the mir17 family, Mirn17/Mir17, with the mRNA gene list mentioned above. This gene list was provided as input to InteractomeBrowser. Figure 2A shows node placement according to cellular compartment. As shown in Figure 2A and 2B this layout is extremely useful to directly focus on genes of interest. Indeed, the nucleus subnetwork contains several regulators (*e.g;* Runx1, Notch1, Hes1 and Xbp1) some of them colored in green, indicating available regulatory interactions for the transcription factor in our database. Figure 2B shows that several genes (Dtx1, Hes1, Il7r and Bcl2) have been previously shown to be under the positive control of Notch1 (these curated informations are derived from LymphTF-DB). According to TargetScan predictions, Mirn17/Mir17 does not seem to target any component of the Notch pathway. In contrast, it is predicted to affect the expression of several transcription regulators including Mycn, Runx1, Smad7 and the H3K27 methyltransferase Ezh1 (by default miRNA are considered as having a negative effect on mRNA and thus edges appear as T-shaped arrows). Moreover, it may also control key components of the cell cycle machinery: Ccnd2 and Cdkn1a. Figure 2D shows informations available from ChIP-X database regarding Mycn. These informations are derived from a ChIP-seq experiment performed on mouse embryonic stem cells by Chen *et al* [38]. Note that according to these results, Mycn could target several transcription factors and thus play a key role during DN3 to DN4 transition. However, in this cellular context such results should be interpreted with caution since no large scale analysis of MYCN targets in DN3 Thymocytes has been reported so far. Among Mycn potential targets, Notch1, is one master switch of early to late thymocyte developmental transition. Thus, one could hypothesize that Mirn17/Mir17 may indirectly affect Notch1 by negatively regulating Mycn. Although, these hypotheses rely on predictions and on the assumption that Mycn binding to Notch promoter is effective in DN3 thymocyte, it clearly underlines the potential of this software in helping researchers to draw new hypotheses using data integration.

Conclusions

InteractomeBrowser and its underlying approach can be compared to the Cerebral (Cell Region-Based Rendering And Layout) plugin of Cytoscape that also combines molecular interactions with a cell-compartment based layout [11].

But there are qualitative differences in the conception of Cerebral and InteractomeBrowser, which make the latest an interesting alternative for exploring networks.

On one hand, Cerebral uses a layered representation of the cell to create a “pathway-like” view of the network of interacting proteins. This layout thus provides a linear organisation of the network. On the other hand, the layout of InteractomeBrowser is based on a schematic view of the entire cell and displays the hierarchical structure of the underlying Gene Ontology subset as nested zones. First, this helps visually separating different parts of the network corresponding to different cellular localisations, as in Cerebral. But this is a more generic visualisation method, in the sense it does not restrict the visual message to an 'input-intermediates-output' mechanism such as in linear pathway diagrams. As a consequence it is suited for a more general study of various types of networks. Moreover, since visual zones correspond to Gene Ontology terms, this layout handles different levels of accuracy in the localisation of proteins: for instance a precisely-annotated protein might be placed in the zone corresponding to “endoplasmic reticulum”, while a less well-annotated can be placed in the more generic, higher level zone “intracellular”.

In Cerebral, each gene product is represented by one instance whose cell compartment may be defined by the user. In contrast, InteractomeBrowser displays, by default, several instances of a given gene product that may be placed in several cell-compartments according to informations provided by the GO Cellular-component ontology. Although this may lead to a more complex graph, it provides a more exhaustive presentation of current knowledge and may draw the attention

of users to unexpected locations of gene products in the cells. The user may choose to delete some of these instances hence selecting *a posteriori* the most representative one.

The main benefit of InteractomeBrowser resides in its direct interaction with the database described in this report. Indeed, it provides a ready-to-use web-based service that requires only few manipulations to retrieve a network of interactions (see video tutorial provided as additional file). . Notably, in addition to physical interactions it offers a unified access to miRNA targets and results from ChIP-Seq experiments derived from CHEA.

Presently, the data sources associated with the InteractomeBrowser plug-in are restricted to human and mouse. Indeed, one of the main objectives of InteractomeBrowser is to help users in creating regulatory maps to study human gene regulatory networks in physiological and pathological conditions. The choice of mouse as an additional organism supported by our database is a natural choice as it is a widely used model of human physiopathology. However, we are already planning to add new organisms in the near future.

As more and more experimentally validated interactions are available, we hope that this tool will prove very useful for researchers.

Availability and requirements

InteractomeBrowser comes as a plugin for TranscriptomeBrowser and is available at : http://tagc.univ-mrs.fr/tbrowser/. Our database is updated on a regular basis. See supplementary material for a video tutorial.

- Project name: InteractomeBrowser
- Project home page: http://tagc.univ-mrs.fr/tbrowser/
- Operating system(s): Platform independent (Java)
- Programming language: Java
- Other requirements: Java > 1.6.X
- License: no license required
- Any restrictions to use by non-academics: none

Competing interests

The authors declare that they have no competing interests

List of abbreviation used

PWM: Position Weight Matrices; GRN: gene regulatory network; GO: Gene Ontology; micro RNA: miRNA; transcription factors (TF); transcription factor binding site (TFBS) ; TranscriptomeBrowser Motif Conservation (TBMC);

Authors contributions

CL, AB, FL, CN, JI and DP conceived the project. CL, AB and FL developed the Java application.

AB, CL and NBP developed the database. DP performed the TFBS analysis. DP, CN and JI supervised the project. DP wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by the Institut National de la Santé et de la Recherche Médicale (Inserm), the Canceropôle PACA and Marseille-Nice Genopole®. Authors acknowledge financial support from the EU ERASysBio Plus ModHeart project. Fabrice Lopez was supported by a fellowship from the EU STREP grant Diamonds and through funding from the IntegraTCell project (ANR, National Research Agency). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors would like to thanks the staff from the TAGC laboratory for helpful discussions and gratefully acknowledge Francois-Xavier Theodule for technical assistance.

Figure legends

Figure 1: Functional enrichment analysis of predicted targets. Annotation terms obtained from various annotation databases were used to performed systematic annotation of all predicted target sets in the mouse. For each pair of term/PWM we computed Fisher's exact test p-value *f*. Each cell of a matrix with terms as row and PWM as column was filled with a score defined as *-log(f)*. (A-I) Representative biclusters found with BiMax are presented.

Figure 2: The InteractomeBrowser plugin. (A) A global and zoom-in view of InteractomeBrowser cell-compartment based layout. Zoom-in view shows some sub-cellular compartments together with node corresponding to gene products. Note that node corresponding to Esr1 appears as green, indicating that regulatory information is available for this gene. (B) Positive interactions (*i.e.;* activations) appear as green edges with normal arrowheads (here Notch1 is the source). (C) Negative interactions (*i.e;* repressions) appear as red edges with T-shaped arrowheads (here Mirn17 is the source). (D) Ambiguous interactions (whose repressive or activating status is unknown) appear as violet arrows with dot arrowheads (here with Mycn as source).

Table

Table 1: A comparison of web tools dedicated to molecular interactions. The table provides an overview of the types of molecular interactions and of the functionalities offered by representative web tools previously published. Informations were obtained from latest articles describing the servers.

		MIR@NT@N	STRING[d]	MotifMap[e]	GeneMANIA	APID[f]	InnateDB	InteractomeBrowser
Database content	Physical protein protein interactions	✗	✓	✓	✓	✓	✓	✓
	Computationally predicted TF targets[a]	✓	✗	✓	✗	✗	✗	✓
	Experimentally observed TF targets[b]	✗	✗	✗	✗	✗	✗	✓
	Predicted miRNA targets	✓	✗	✗	✗	✗	✗	✓
	Regulatory interactions from literature	✗	✓	✗	✗	✗	✗	✓
	Biological pathways	✗	✓	✗	✓	✗	✗	✗
	Inferred functional interactions[c]	✗	✓	✗	✓	✗	✗	✗
Build-in graph visualizer	Batch query	✓	✓	✗	✓	✗	✗	✓
	add/remove/hide interactors and interactions	✗	✗	✗	✗	✓	✗	✓
	Movable nodes	✗	✓	ND	✓	✓	✓	✓
	Compartment-based layout	✗	✗	✗	✗	✗	✓	✓

[a] Refers to bioinformatic prediction of TFBSs using PWMs.

[b] Refers to results from large-scale experimental methods that profile the binding of TFs to DNA at the genome-wide level (*e.g.*; ChIP-Seq, ChIP-chip, …).

[c] Refers to computational methods that aggregate various informations (*e.g.*; expression, genomic distance, conservation) to infer functional interactions.

[d] Search Tool for the Retrieval of Interacting Genes/Proteins

[e] MotifMap visualizer was not available during our tests. Informations related to the visualizer were obtained from documentation.

[f] Agile Protein Interaction DataAnalyzer

Description of additional data files

- File name: Fig.S1.pdf
    - File format : pdf
    - Title: “Number of predicted motifs versus GC content of PWMs”
    - Description of data: Each point corresponds to the results obtained using one PWM on mouse genome. The name of a representative transcription factor for each PWM is displayed together with the PWM identifier (informations are separated using a pipe character). The size of the point is correlated with info content of the corresponding matrix).

- File name: Fig.S2.pdf
    - File format : pdf
    - Title: “Summary of functional enrichment analysis using ClueGO cytoscape plugin”
    - Description of data: We estimated the number of predicted regulators for each gene of the human genome by computing the number of non-redundant position-specific motifs associated with each genes. Genes in the top 1% regards to the number of regulators were used as input for the ClueGO plugin.

- File name: Fig.S3.pdf
    - File format : pdf
    - Title: “Summary of functional enrichment analysis using ClueGO cytoscape plugin”
    - Description of data: We estimated the number of predicted regulators for each gene of the mouse genome by computing the number of non- redundant position-specific motifs

associated with each genes. Genes in the top 1% regards to the number of regulators were used as input for the ClueGO plugin.

- File name: Fig.S4.pdf
  - File format : pdf
  - Title: “Subset of Gene Ontology used for the cell compartment-based layout ”

    Description of data: Hierarchical structure of the subset of Gene Ontology used in InteractomeBrowser for the cell compartment-based layout. Colors highlight the main compartments.

- File name: TBMC.mm.bed
  - File format : bed
  - Title: “TFBS predictions in the mouse genome”

    Description of data: A bed file containing TFBS predictions in the mouse genome. 1 - chrom - The name of the chromosome. Fields contain the following informations: chromStart - The starting position of the feature in the chromosome; chromEnd - The ending position of the feature in the chromosome; name - PWM identifier and representative names; score - A score for the PWM hit; strand - Defines the strand - either '+' or '-'; gene id - The gene id of the target gene; geneSymbol- The genesymbol of the target gene.

- File name: TBMC.hs.bed
  - File format : bed

- Title: “TFBS predictions in the human genome”

  Description of data: A bed file containing TFBS predictions in the human genome. 1 - chrom - The name of the chromosome. Fields contain the following informations: chromStart - The starting position of the feature in the chromosome; chromEnd - The ending position of the feature in the chromosome; name - PWM identifier and representative names; score - A score for the PWM hit; strand - Defines the strand - either '+' or '-'; gene id - The gene id of the target gene; geneSymbol- The genesymbol of the target gene.

- File name: Video.tutorial.doc
  - File format : doc
  - Title: “InteractomeBrowser functionalities”

    Description of data: Contains a web link to a screencast showing basic use of InteractomeBrowser plugin.

References

1. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A: **NCBI GEO: archive for functional genomics data sets--10 years on**. *Nucleic Acids Res* 2011, **39**:D1005-1010.

2. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals**. *Nature* 2005, **434**:338-345.

3. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK: **Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data**. *Genome Res* 2011, **21**:447-455.

4. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D, Gavin A-C: **Visualization of omics data for systems biology**. *Nat. Methods* 2010, **7**:S56-68.

5. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**:2498-2504.

6. Le Béchec A, Portales-Casamar E, Vetter G, Moes M, Zindy P-J, Saumet A, Arenillas D, Theillet C, Wasserman WW, Lecellier C-H, Friederich E: **MIR@NT@N: a framework integrating transcription factors, microRNAs and their targets to identify sub-network motifs in a meta-regulation network model**. *BMC Bioinformatics* 2011, **12**:67.

7. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C: **The STRING database in 2011: functional**

interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 2011, **39**:D561-568.

8. Xie X, Rigor P, Baldi P: **MotifMap: a human genome-wide map of candidate regulatory motif sites**. *Bioinformatics* 2009, **25**:167-174.

9. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD, Morris Q: **The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function**. *Nucleic Acids Res* 2010, **38**:W214-220.

10. Hernandez-Toro J, Prieto C, De las Rivas J: **APID2NET: unified interactome graphic analyzer**. *Bioinformatics* 2007, **23**:2495-2497.

11. Barsky A, Gardy JL, Hancock REW, Munzner T: **Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation**. *Bioinformatics* 2007, **23**:1040 -1042.

12. Wingender E, Dietze P, Karas H, Knüppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites**. *Nucleic Acids Res* 1996, **24**:238-241.

13. A. Sandelin: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles**. *Nucleic Acids Research* 2004, **32**:91D-94.

14. Newburger DE, Bulyk ML: **UniPROBE: an online database of protein binding microarray data on protein-DNA interactions**. *Nucleic Acids Res* 2009, **37**:D77-82.

15. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A: **ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments**. *Bioinformatics* 2010, **26**:2438-2444.

16. Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, Griffith M, Gallo SM, Giardine B, Hooghe B, Van Loo P, Blanco E, Ticoll A, Lithwick S, Portales-Casamar E, Donaldson IJ, Robertson G, Wadelius C, De Bleser P, Vlieghe D, Halfon MS, Wasserman W, Hardison R, Bergman CM, Jones SJM: **ORegAnno: an open-access community-driven resource for regulatory annotation**. *Nucleic Acids Res* 2008, **36**:D107-113.

17. Childress PJ, Fletcher RL, Perumal NB: **LymphTF-DB: a database of transcription factors involved in lymphocyte development**. *Genes Immun* 2007, **8**:360-365.

18. Friedman RC, Farh KK-H, Burge CB, Bartel DP: **Most mammalian mRNAs are conserved targets of microRNAs**. *Genome Research* 2009, **19**:92 -105.

19. Lachmann A, Ma'ayan A: **KEA: kinase enrichment analysis**. *Bioinformatics* 2009, **25**:684 -686.

20. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, Kerssemakers J, Leroy C, Menden M, Michaut M, Montecchi-Palazzi L, Neuhauser SN, Orchard S, Perreau V, Roechert B, van Eijk K, Hermjakob H: **The IntAct molecular interaction database in 2010**. *Nucleic Acids Research* 2009.

21. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A: **Human Protein Reference Database--2009 update**. *Nucleic Acids Res* 2009, **37**:D767-772.

22. Lopez F, Textoris J, Bergon A, Didier G, Remy E, Granjeaud S, Imbert J, Nguyen C, Puthier D: **TranscriptomeBrowser: a powerful and flexible toolbox to explore productively the**

transcriptional landscape of the Gene Expression Omnibus database. *PLoS ONE* 2008, **3**:e4001.

23. Naldi A, Berenguier D, Fauré A, Lopez F, Thieffry D, Chaouiya C: **Logical modelling of regulatory networks with GINsim 2.3**. *BioSystems* 2009, **97**:134-139.

24. Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Pohl A, Malladi VS, Li CH, Learned K, Kirkup V, Hsu F, Harte RA, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, James Kent W: **The UCSC Genome Browser database: extensions and updates 2011**. *Nucleic Acids Research* 2011.

25. Stormo GD: **DNA binding sites: representation and discovery**. *Bioinformatics* 2000, **16**:16-23.

26. Thomas-Chollier M, Sand O, Turatsinze J-V, Janky R, Defrance M, Vervisch E, Brohée S, van Helden J: **RSAT: regulatory sequence analysis tools**. *Nucleic Acids Res* 2008, **36**:W119-127.

27. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR: **Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs**. *Cell* 2004, **116**:499-509.

28. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman W-H, Pagès F, Trajanoski Z, Galon J: **ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks**. *Bioinformatics* 2009, **25**:1091-1093.

29. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D:

Ultraconserved elements in the human genome. *Science* 2004, **304**:1321-1325.

30. Gerstein MB, Lu ZJ, Van Nostrand EL, et al.: **Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project**. *Science* 2010, **330**:1775 -1787.

31. The modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S, Arshinoff BI, Ay F, Meyer PE, Robine N, Washington NL, Di Stefano L, Berezikov E, Brown CD, Candeias R, Carlson JW, Carr A, Jungreis I, Marbach D, Sealfon R, Tolstorukov MY, Will S, Alekseyenko AA, Artieri C, Booth BW, Brooks AN, Dai Q, Davis CA, Duff MO, Feng X, Gorchakov AA, Gu T, Henikoff JG, Kapranov P, Li R, MacAlpine HK, Malone J, Minoda A, Nordman J, Okamura K, Perry M, Powell SK, Riddle NC, Sakai A, Samsonova A, Sandler JE, Schwartz YB, Sher N, Spokony R, Sturgill D, van Baren M, Wan KH, Yang L, Yu C, Feingold E, Good P, Guyer M, Lowdon R, Ahmad K, Andrews J, Berger B, Brenner SE, Brent MR, Cherbas L, Elgin SCR, Gingeras TR, Grossman R, Hoskins RA, Kaufman TC, Kent W, Kuroda MI, Orr-Weaver T, Perrimon N, Pirrotta V, Posakony JW, Ren B, Russell S, Cherbas P, Graveley BR, Lewis S, Micklem G, Oliver B, Park PJ, Celniker SE, Henikoff S, Karpen GH, Lai EC, MacAlpine DM, Stein LD, White KP, Kellis M: **Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE**. *Science* 2010, **330**:1787 -1797.

32. Bader GD, Cary MP, Sander C: **Pathguide: a pathway resource list**. *Nucleic Acids Res* 2006, **34**:D504-506.

33. Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E: **A systematic comparison and evaluation of biclustering methods for gene expression data**. *Bioinformatics* 2006, **22**:1122-1129.

34. Nie L, Xu M, Vladimirova A, Sun X-H: **Notch-induced E2A ubiquitination and degradation**

are controlled by MAP kinase activities. *EMBO J* 2003, **22**:5780-5792.

35. Aranburu A, Carlsson R, Persson C, Leanderson T: **Transcription factor AP-4 is a ligand for immunoglobulin-kappa promoter E-box elements.** *Biochem J* 2001, **354**:431-438.

36. Painter MW, Davis S, Hardy RR, Mathis D, Benoist C: **Transcriptomes of the B and T lineages compared by multiplatform microarray profiling**. *J. Immunol* 2011, **186**:3047-3057.

37. Neilson JR, Zheng GXY, Burge CB, Sharp PA: **Dynamic regulation of miRNA expression in ordered stages of cellular development**. *Genes & Development* 2007, **21**:578 -589.

38. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh Y-H, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung W-K, Clarke ND, Wei C-L, Ng H-H: **Integration of external signaling pathways with the core transcriptional network in embryonic stem cells**. *Cell* 2008, **133**:1106-1117.

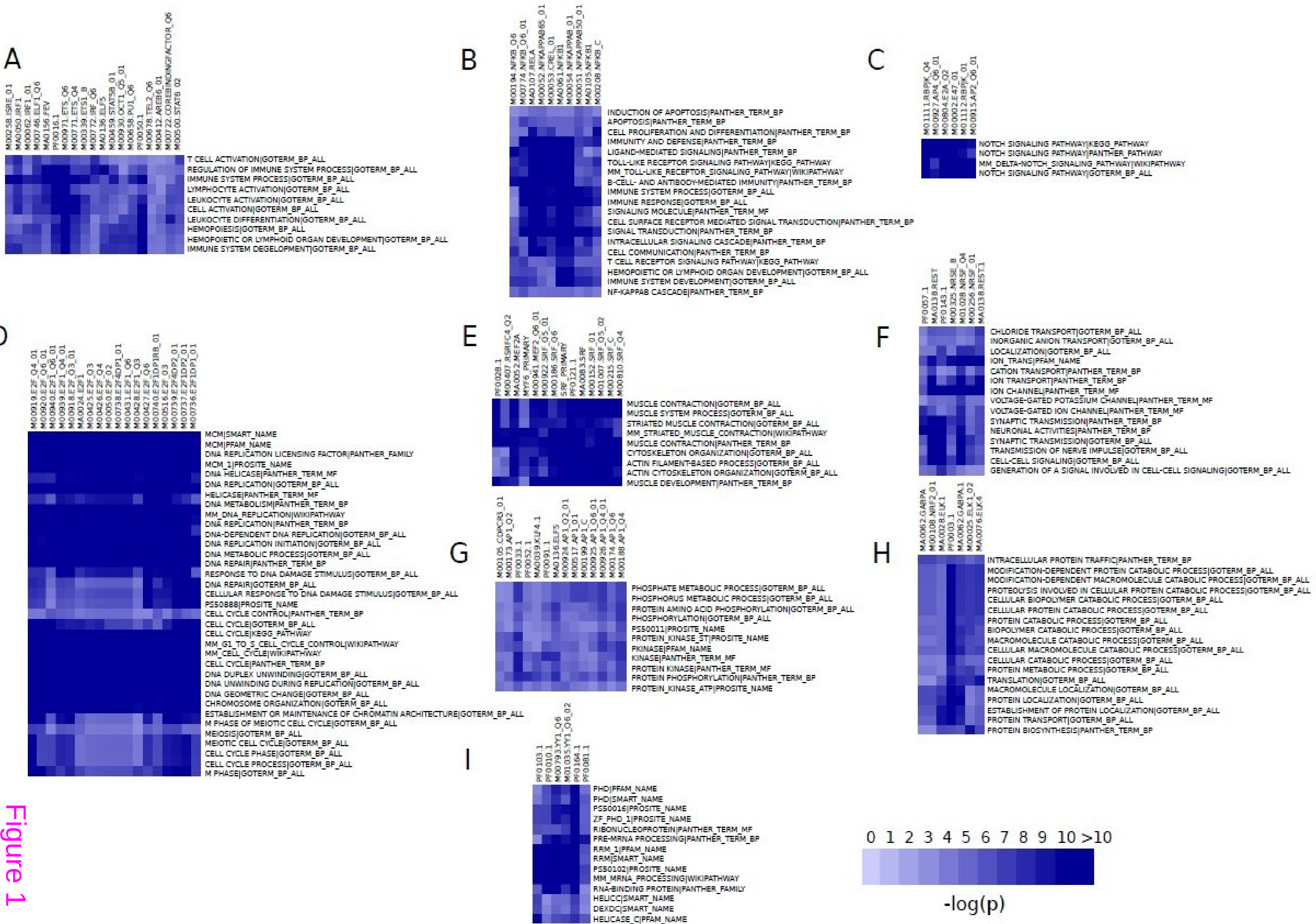

A
M00258.ISRE_01
MA0050.IRF1
M00062.IRF1_01
M00746.ELF1_Q6
MA0156.FEV
PF0016.1
M00971.ETS_Q6
M00771.ETS_Q4
M00339.ETS1_B
M00772.IRF_Q6
MA0136.ELF5
M00459.STAT5B_01
M00930.OCT1_Q5_01
M00658.PU1_Q6
PF0050.1
M00678.TEL2_Q6
M00412.AREB6_01
M00722.COREBINDINGFACTOR_Q6
M00500.STAT6_02
T CELL ACTIVATION|GOTERM_BP_ALL
REGULATION OF IMMUNE SYSTEM PROCESS|GOTERM_BP_ALL
IMMUNE SYSTEM PROCESS|GOTERM_BP_ALL
LYMPHOCYTE ACTIVATION|GOTERM_BP_ALL
LEUKOCYTE ACTIVATION|GOTERM_BP_ALL
CELL ACTIVATION|GOTERM_BP_ALL
LEUKOCYTE DIFFERENTIATION|GOTERM_BP_ALL
HEMOPOIESIS|GOTERM_BP_ALL
HEMOPOIETIC OR LYMPHOID ORGAN DEVELOPMENT|GOTERM_BP_ALL
IMMUNE SYSTEM DEGELOPMENT|GOTERM_BP_ALL
B
M00194.NFKB_Q6
M00774.NFKB_Q6_01
MA0107.RELA
M00052.NFKAPPAB65_01
M00053.CREL_01
MA0061.NFKB1
M00054.NFKAPPAB_01
MA0105.NFKB1
M00051.NFKAPPAB50_01
M00208.NFKB_C
INDUCTION OF APOPTOSIS|PANTHER_TERM_BP
APOPTOSIS|PANTHER_TERM_BP
CELL PROLIFERATION AND DIFFERENTIATION|PANTHER_TERM_BP
IMMUNITY AND DEFENSE|PANTHER_TERM_BP
LIGAND-MEDIATED SIGNALING|PANTHER_TERM_BP
TOLL-LIKE RECEPTOR SIGNALING PATHWAY|KEGG_PATHWAY
MM_TOLL-LIKE_RECEPTOR_SIGNALING_PATHWAY|WIKIPATHWAY
B-CELL- AND ANTIBODY-MEDIATED IMMUNITY|PANTHER_TERM_BP
IMMUNE SYSTEM PROCESS|GOTERM_BP_ALL
IMMUNE RESPONSE|GOTERM_BP_ALL
SIGNALING MOLECULE|PANTHER_TERM_MF
CELL SURFACE RECEPTOR MEDIATED SIGNAL TRANSDUCTION|PANTHER_TERM_BP
SIGNAL TRANSDUCTION|PANTHER_TERM_BP
INTRACELLULAR SIGNALING CASCADE|PANTHER_TERM_BP
CELL COMMUNICATION|PANTHER_TERM_BP
T CELL RECEPTOR SIGNALING PATHWAY|KEGG_PATHWAY
HEMOPOIETIC OR LYMPHOID ORGAN DEVELOPMENT|GOTERM_BP_ALL
IMMUNE SYSTEM DEVELOPMENT|GOTERM_BP_ALL
NF-KAPPAB CASCADE|PANTHER_TERM_BP
C
M01111.RBPJK_Q4
M00927.AP4_Q6_01
M00804.E2A_Q2
M00002.E47_01
M01112.RBPJK_01
M00915.AP2_Q6_01
NOTCH SIGNALING PATHWAY|KEGG_PATHWAY
NOTCH SIGNALING PATHWAY|PANTHER_PATHWAY
MM_DELTA-NOTCH_SIGNALING_PATHWAY|WIKIPATHWAY
NOTCH SIGNALING PATHWAY|GOTERM_BP_ALL
D
M00919.E2F_Q4_01
M00920.E2F_Q6_01
M00940.E2F1_Q6_01
M00939.E2F_Q4_01
M00918.E2F_Q3_01
MA0024.E2F1
M00425.E2F_Q3
M00426.E2F_Q4
M00050.E2F_02
M00738.E2F4DP1_01
M00431.E2F1_Q6
M00428.E2F1_Q3
M00427.E2F_Q6
M00740.E2F1DP1RB_01
M00516.E2F_03
M00739.E2F4DP2_01
M00737.E2F1DP2_01
M00736.E2F1DP1_01
MCM|SMART_NAME
MCM|PFAM_NAME
DNA REPLICATION LICENSING FACTOR|PANTHER_FAMILY
MCM_1|PROSITE_NAME
DNA HELICASE|PANTHER_TERM_MF
DNA REPLICATION|GOTERM_BP_ALL
HELICASE|PANTHER_TERM_MF
DNA METABOLISM|PANTHER_TERM_BP
MM_DNA_REPLICATION|WIKIPATHWAY
DNA REPLICATION|PANTHER_TERM_BP
DNA-DEPENDENT DNA REPLICATION|GOTERM_BP_ALL
DNA REPLICATION INITIATION|GOTERM_BP_ALL
DNA METABOLIC PROCESS|GOTERM_BP_ALL
DNA REPAIR|PANTHER_TERM_BP
RESPONSE TO DNA DAMAGE STIMULUS|GOTERM_BP_ALL
DNA REPAIR|GOTERM_BP_ALL
CELLULAR RESPONSE TO DNA DAMAGE STIMULUS|GOTERM_BP_ALL
PS50888|PROSITE_NAME
CELL CYCLE CONTROL|PANTHER_TERM_BP
CELL CYCLE|GOTERM_BP_ALL
CELL CYCLE|KEGG_PATHWAY
MM_G1_TO_S_CELL_CYCLE_CONTROL|WIKIPATHWAY
MM_CELL_CYCLE|WIKIPATHWAY
CELL CYCLE|PANTHER_TERM_BP
DNA DUPLEX UNWINDING|GOTERM_BP_ALL
DNA UNWINDING DURING REPLICATION|GOTERM_BP_ALL
DNA GEOMETRIC CHANGE|GOTERM_BP_ALL
CHROMOSOME ORGANIZATION|GOTERM_BP_ALL
ESTABLISHMENT OR MAINTENANCE OF CHROMATIN ARCHITECTURE|GOTERM_BP_ALL
M PHASE OF MEIOTIC CELL CYCLE|GOTERM_BP_ALL
MEIOSIS|GOTERM_BP_ALL
MEIOTIC CELL CYCLE|GOTERM_BP_ALL
CELL CYCLE PHASE|GOTERM_BP_ALL
CELL CYCLE PROCESS|GOTERM_BP_ALL
M PHASE|GOTERM_BP_ALL
E
PF0028.1
M00407.RSRFC4_Q2
MA0052.MEF2A
MYF6_PRIMARY
M00941.MEF2_Q6_01
M00922.SRF_Q5_01
M00186.SRF_Q6
SRF_PRIMARY
PF0121.1
MA0083.SRF
M00152.SRF_01
M01007.SRF_Q5_02
M00215.SRF_C
M00810.SRF_Q4
MUSCLE CONTRACTION|GOTERM_BP_ALL
MUSCLE SYSTEM PROCESS|GOTERM_BP_ALL
STRIATED MUSCLE CONTRACTION|GOTERM_BP_ALL
MM_STRIATED_MUSCLE_CONTRACTION|WIKIPATHWAY
MUSCLE CONTRACTION|PANTHER_TERM_BP
CYTOSKELETON ORGANIZATION|GOTERM_BP_ALL
ACTIN FILAMENT-BASED PROCESS|GOTERM_BP_ALL
ACTIN CYTOSKELETON ORGANIZATION|GOTERM_BP_ALL
MUSCLE DEVELOPMENT|PANTHER_TERM_BP
F
PF0057.1
MA0138.REST
PF0143.1
M00325.NRSE_B
M01028.NRSF_Q4
M00256.NRSF_01
MA0138.REST.1
CHLORIDE TRANSPORT|GOTERM_BP_ALL
INORGANIC ANION TRANSPORT|GOTERM_BP_ALL
LOCALIZATION|GOTERM_BP_ALL
ION_TRANS|PFAM_NAME
CATION TRANSPORT|PANTHER_TERM_BP
ION TRANSPORT|PANTHER_TERM_BP
ION CHANNEL|PANTHER_TERM_MF
VOLTAGE-GATED POTASSIUM CHANNEL|PANTHER_TERM_MF
VOLTAGE-GATED ION CHANNEL|PANTHER_TERM_MF
SYNAPTIC TRANSMISSION|PANTHER_TERM_BP
NEURONAL ACTIVITIES|PANTHER_TERM_BP
SYNAPTIC TRANSMISSION|GOTERM_BP_ALL
TRANSMISSION OF NERVE IMPULSE|GOTERM_BP_ALL
CELL-CELL SIGNALING|GOTERM_BP_ALL
GENERATION OF A SIGNAL INVOLVED IN CELL-CELL SIGNALING|GOTERM_BP_ALL
G
M00105.CDPCR3_01
M00173.AP1_Q2
PF0033.1
PF0052.1
MA0039.KLF4.1
PF0091.1
MA0136.ELF5
M00924.AP1_Q2_01
M00517.AP1_01
M00199.AP1_C
M00925.AP1_Q6_01
M00926.AP1_Q4_01
M00174.AP1_Q6
M00188.AP1_Q4
PHOSPHATE METABOLIC PROCESS|GOTERM_BP_ALL
PHOSPHORUS METABOLIC PROCESS|GOTERM_BP_ALL
PROTEIN AMINO ACID PHOSPHORYLATION|GOTERM_BP_ALL
PHOSPHORYLATION|GOTERM_BP_ALL
PS50011|PROSITE_NAME
PROTEIN_KINASE_ST|PROSITE_NAME
PKINASE|PFAM_NAME
KINASE|PANTHER_TERM_MF
PROTEIN KINASE|PANTHER_TERM_MF
PROTEIN PHOSPHORYLATION|PANTHER_TERM_BP
PROTEIN_KINASE_ATP|PROSITE_NAME
H
MA0062.GABPA
M00108.NRF2_01
MA0028.ELK1
PF0003.1
MA0062.GABPA.1
M00025.ELK1_02
MA0076.ELK4
INTRACELLULAR PROTEIN TRAFFIC|PANTHER_TERM_BP
MODIFICATION-DEPENDENT PROTEIN CATABOLIC PROCESS|GOTERM_BP_ALL
MODIFICATION-DEPENDENT MACROMOLECULE CATABOLIC PROCESS|GOTERM_BP_ALL
PROTEOLYSIS INVOLVED IN CELLULAR PROTEIN CATABOLIC PROCESS|GOTERM_BP_ALL
CELLULAR BIOPOLYMER CATABOLIC PROCESS|GOTERM_BP_ALL
CELLULAR PROTEIN CATABOLIC PROCESS|GOTERM_BP_ALL
PROTEIN CATABOLIC PROCESS|GOTERM_BP_ALL
BIOPOLYMER CATABOLIC PROCESS|GOTERM_BP_ALL
MACROMOLECULE CATABOLIC PROCESS|GOTERM_BP_ALL
CELLULAR MACROMOLECULE CATABOLIC PROCESS|GOTERM_BP_ALL
CELLULAR CATABOLIC PROCESS|GOTERM_BP_ALL
PROTEIN METABOLIC PROCESS|GOTERM_BP_ALL
TRANSLATION|GOTERM_BP_ALL
MACROMOLECULE LOCALIZATION|GOTERM_BP_ALL
PROTEIN LOCALIZATION|GOTERM_BP_ALL
ESTABLISHMENT OF PROTEIN LOCALIZATION|GOTERM_BP_ALL
PROTEIN TRANSPORT|GOTERM_BP_ALL
PROTEIN BIOSYNTHESIS|PANTHER_TERM_BP
I
PF0103.1
PF0010.1
M00793.YY1_Q6
M01035.YY1_Q6_02
PF0164.1
PF0081.1
PHD|PFAM_NAME
PHD|SMART_NAME
PS50016|PROSITE_NAME
ZF_PHD_1|PROSITE_NAME
RIBONUCLEOPROTEIN|PANTHER_TERM_MF
PRE-MRNA PROCESSING|PANTHER_TERM_BP
RRM_1|PFAM_NAME
RRM|SMART_NAME
PS50102|PROSITE_NAME
MM_MRNA_PROCESSING|WIKIPATHWAY
RNA-BINDING PROTEIN|PANTHER_FAMILY
HELICC|SMART_NAME
DEXDC|SMART_NAME
HELICASE_C|PFAM_NAME
0 1 2 3 4 5 6 7 8 9 10 >10
-log(p)

Figure 1

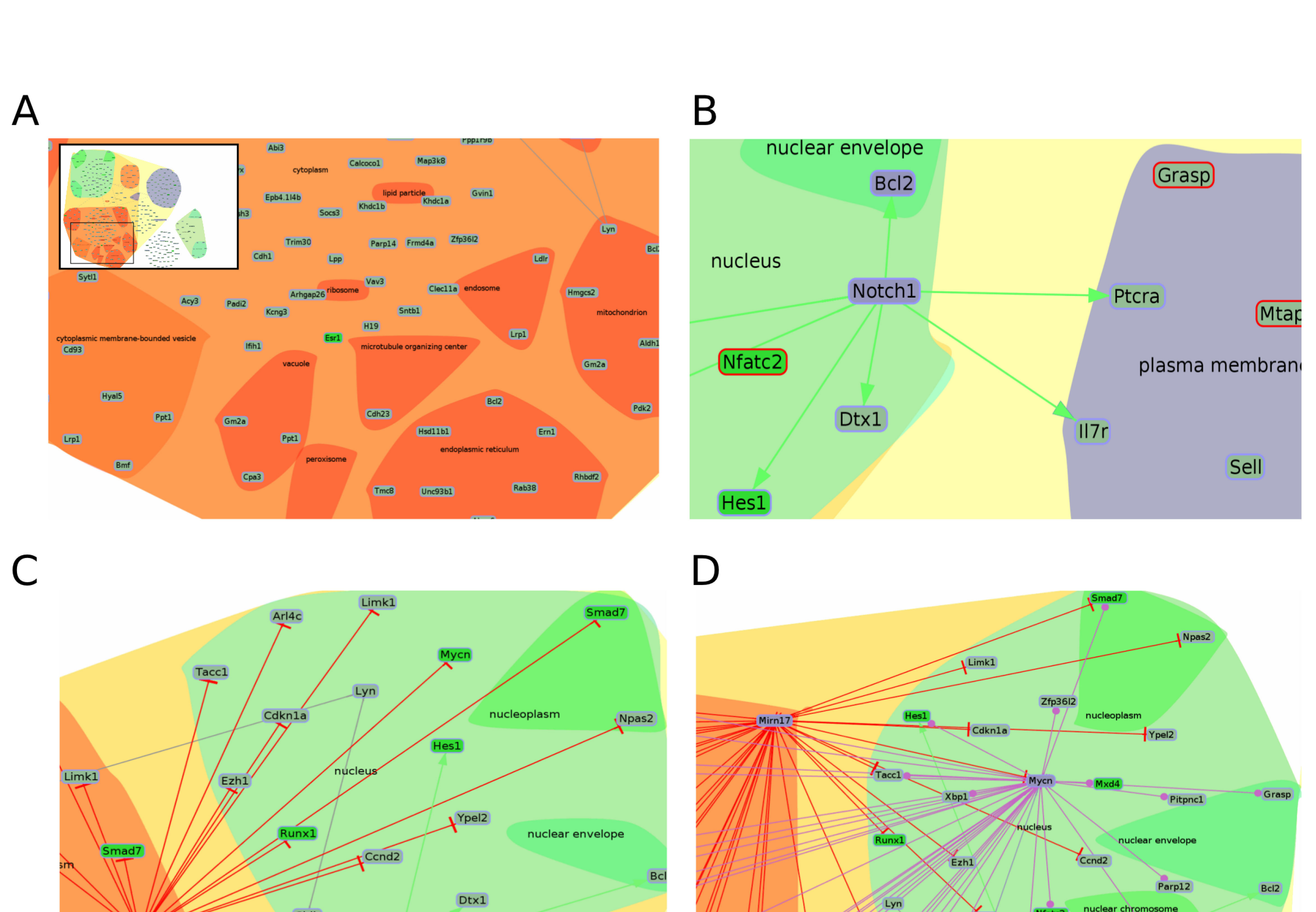

A
B
C
D
nuclear envelope
Bcl2
nucleus
Notch1
Ptcra
Grasp
Nfatc2
Dtx1
Il7r
plasma membrane
Hes1
Sell
Mirn17
Smad7
Mycn
nucleoplasm
Npas2
Hes1
Ypel2
Ccnd2
Dtx1
Runx1
Ezh1
Cdkn1a
Tacc1
Arl4c
Limk1
Lyn
Cblb
Mtap7
Mxd4
Pitpnc1
Parp12
nuclear chromosome
Smarca4
Zfp36l2
Xbp1
cytoplasm
mitochondrion
endoplasmic reticulum
vacuole
peroxisome
endosome
ribosome
microtubule organizing center
cytoplasmic membrane-bounded vesicle
lipid particle

Figure 2

Additional files provided with this submission:

Additional file 1: Fig.S1.pdf, 20K
http://www.biomedcentral.com/imedia/1502059913646762/supp1.pdf
Additional file 2: Fig.S2.pdf, 10K
http://www.biomedcentral.com/imedia/3065379596467629/supp2.pdf
Additional file 3: Fig.S3.pdf, 21K
http://www.biomedcentral.com/imedia/1156508138646762/supp3.pdf
Additional file 4: Fig.S4.pdf, 18K
http://www.biomedcentral.com/imedia/2672681166467629/supp4.pdf
Additional file 5: TBMC.mm.bed, 5802K
http://www.biomedcentral.com/imedia/1789905813646762/supp5.bed
Additional file 6: TBMC.hs.bed, 4849K
http://www.biomedcentral.com/imedia/1455949074646762/supp6.bed
Additional file 7: Video.tutorial.doc, 9K
http://www.biomedcentral.com/imedia/1319895762646762/supp7.doc

4.6.3 Les cartes transcriptionnelles pour le plugin TBMap

J'ai créé des procédures stockées permettant de générer des cartes transcriptionnelles pour diverses espèces à partir de cette nouvelle base de données où les informations ne sont plus organisées de la même manière.

J'ai également créé un script qui permet de générer une carte transcriptionnelle à partir d'une liste de gènes et pour toutes les signatures possédant au minimum un gène de cette liste. Notre base de données contenant maintenant beaucoup plus d'espèces, j'ai également modifié le script original pour accepter en entrée les homologeneID, permettant ainsi d'avoir une carte comprenant plusieurs espèces proches. Ces cartes peuvent également être visualiser à l'aide d'outils tels que Treeview et TMeV, logiciels pour l'analyse de données de puces à ADN.

4.7 Accès programmé à la base de données de TBrowser

Afin de permettre un accès à notre base de données, par des outils de programmation aux utilisateurs experts, nous avons développé des services web et une librairie R y accédant, ceci dans le but de permettre l'exploitation des données, mais également leur génération.

4.7.1 Développement de services web

Deux versions de service web ont été mises au point durant ma thèse dans le but de rendre accessible les données contenues dans notre première base de données puis une fois réalisée, celles de notre nouvelle base de données de signatures transcriptionnelles.

(1) La première est sous forme d'une procédure servlet java qui interagit avec la première version de la base de données. La documentation des fonctions implémentées est accessible à l'adresse http://tagc.univ-mrs.fr/tbrowser/ws/ et est utilisable tel que : http ://tagc.univ-mrs.fr/tB/TBWS/servlets/TBWS ?type=field&request=value où field peut correspondre à : « gene », « probe », « GSE », « GPL », « signature » et « annotation » et value à une requête booléenne, avec des opérateurs logiques ou « | », et « * », ou une requête par liste avec les gènes séparés par des « , ». Il est à noter qu'ici on utilise un « * » au lieu de « & » comme opérateur logique car « & » est utilisé pour séparer les différents paramètres de la requête du service web.

(2) La seconde et plus récente version correspond à un service web en java basé sur SOAP/WSDL, tournant sur un serveur Tomcat / Apache Axis2™et accédant aux données contenues dans la nouvelle base de données. Ce type de service web est de plus en plus utilisé en bioinformatique, et est retrouvé pour des outils tels que RSATools ou bien encore des bases de données comme Kegg. La documentation de notre service web est accessible à l'adresse : http://tagc.univ-mrs.fr/services/TBService?wsdl Les mêmes requêtes que pour la précédente version ont été développées, mais ce service web n'interroge que la dernière version de la base de données via l'appel de procédures stockées.
Cette dernière peut être intégrée à des workflows de type Taverna, ou à des outils tels que

Cytoscape, ce qui permet l'utilisation de notre base de données à travers d'autres outils.

4.7.2 Implémentation d'une librairie R/Bioconductor : RTools4TB

Cette librairie est constituée d'un ensemble d'objets et de fonctions codées sous R, et d'un programme écrit en langage C permettant :
(1) d'interroger la nouvelle base de données via l'utilisation du service web ;
(2) d'extraire des signatures transcriptionnelles à l'aide de l'algorithme DBF-MCL à partir d'une matrice d'expression.

Le programme C appelé par le code R réalise la première partie de l'algorithme et génère les données pour MCL. Cette librairie fait également appel à d'autres librairies R telles que Biobase, limma, methods, XML, Rcurl et SSOAP. Enfin le code R appelle les programmes mcl et cluster grâce à des commandes système.

Il est ainsi possible de questionner la base de données afin d'obtenir des informations sur une expérience, une plateforme de puce à ADN, une signature à l'aide de la fonction getTBInfo, mais également de récupérer une liste de signatures correspondant au résultat d'une requête (booléenne ou par liste) par la fonction getSignatures. Il est donc possible de concevoir des scripts R pour programmer l'extraction des données et les analyser plus facilement sans passer par l'interface graphique de TBrowser.

Cette librairie permet également l'extraction de signatures transcriptionnelles à l'aide de l'algorithme DBF-MCL, par la fonction DBFMCL. Celle-ci permet d'avoir accès aux différents paramètres de l'algorithme tel que l'inflation, le nombre k de plus proches voisins, le nombre de randomisations, la valeur de FDR (False Discovery Rate) utilisée. Cette fonction utilise en entrée une matrice d'expression et renvoie un objet de type S4 « DBFMCLresult » contenant les paramètres de l'algorithme, la matrice d'expression de chaque signature, le nombre de sondes. Comme dans la version initiale, seules les signatures de plus de 10 sondes sont conservées. L'utilisateur peut ainsi utiliser notre algorithme sur son jeu de données et créer de nouvelles signatures.

Comme pour la précédente librairie R, des fiches d'aide ont été créées pour chaque fonction ou classe d'objet créée ainsi qu'un manuel d'utilisation (voir ci-après).

Cette librairie R a été intégrée aux librairies présentes dans Bioconductor à partir de la version 2.5. La page web de la librairie sur le site de Bioconductor est accessible à l'adresse : http://www.bioconductor.org/packages/2.8/bioc/html/RTools4TB.html (version 2.8 de Bioconductor actuellement). Une nouvelle version de la librairie utilisant le service web SOAP/WSDL est en cours de préparation et sera bientôt mise en ligne sur le site de Bioconductor, via un système de gestion de version de type SVN. Celle-ci permettra également d'annoter une liste de gènes à partir des données d'annotation contenues dans notre base de données.

Un résumé de l'utilisation de la librairie RTools4TB et de son utilisation est présenté dans le manuel d'utilisation téléchargeable avec la librairie (voir Annexe B).

Afin de permettre un développement conjoint par tous les développeurs du projet, et également l'archivage et la maintenance de celui-ci, un système de gestion de version de type SVN a été mis en place. Un article résumant l'avancé du projet (Figure 4.7 et Table 4.1) depuis sa publication initiale en 2008 est en cours de préparation.

4.8 Conclusions et perspectives

Utilisation de TBrowser

Une des manière de mesurer l'impact de l'utilisation de TBrowser sur la commauté scientifique est d'étudier son utilisation. L'article de TBrowser a été vu 2396 fois à partir du site de PLoS ONE au 11 janvier 2011.

TBrowser peut être utilisé comme outils de visualisation grâce au plugin IntractomeBrowser comme ça a été déjà le cas ([Textoris *et al.*, 2010]) ou comme base de données de signature transcriptionnelle au même titre que MsigDB (Molecular Signature Data Base).
Les données de coexpression de gènes de TBrowser sont ainsi utilisées par l'outil PredictSearch®développé par la société Prédiguard (Marseille). PredictSearch est un outil commercial permettant la construction d'un réseau génique à partir d'une sélection drastique de gènes ([Baron *et al.*, 2011]).

De même pour la librairie R RTools4TB, Bioconductor génère des statistiques de téléchargement des librairies, uniquement sur ceux des 12 derniers mois. Bien que le téléchargement de la librairie ait baissé, du fait qu'elle n'est toujours pas publié à ce jour, elle compte cependant 952 téléchargements depuis février 2011.

Prioritization de gènes

Il est également possible d'utiliser des approches telles que la prioritization de gènes afin de mettre en évidence des gènes intéressant parmis les nombreux gènes sélectionnés lors de l'analyse de données de puces à ADN.

En effet, l'identification de gènes clés impliqués dans une maladie reste un défi important de la recherche médicale. Plusieurs approches de prioritization de gènes ont été développées, comme Endeavour [Aerts *et al.*, 2006]. Les données provenant de multiples sources hétérogènes sont intégrées : l'expression des gènes, leur annotation, la littérature, la structure en domaine des protéines et leur interaction, etc Toutes ces données sont ainsi utilisés séparemment afin d'ordonner sur la base de similarité, une liste de gène de référence (spécifique de la maladie étudiée) à une liste de gènes. Puis le rang obtenu pour chaque type d'information est fusionné pour chaque gène, afin d'obtenir un classement globale.

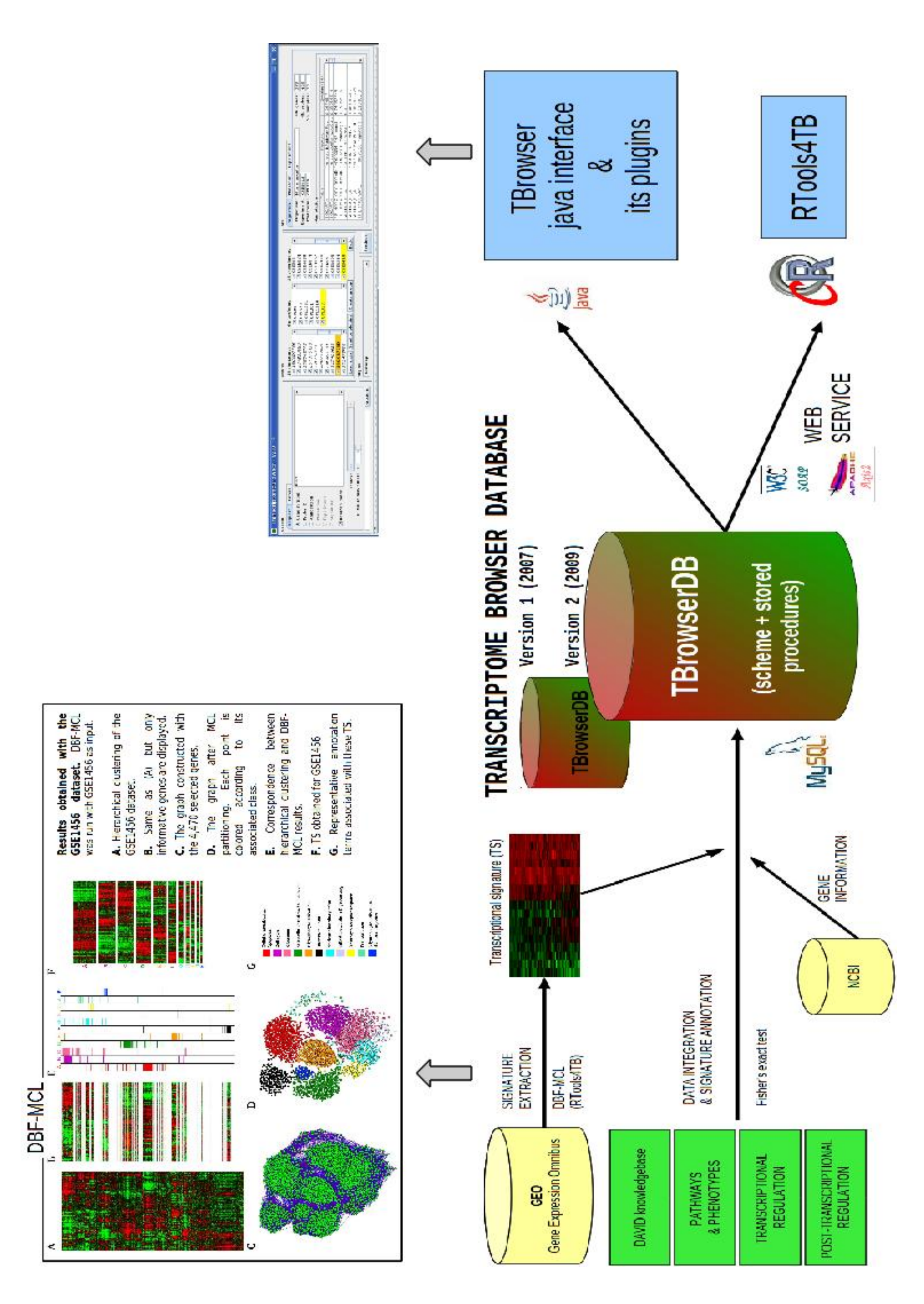


FIGURE 4.7 – Résumé de l'avancement du projet avec le développement d'une nouvelle base de données, d'une librairie R, de services web, et l'intégration de nouvelles données

Feature	TranscriptomeBrowser 2008	Added in TranscriptomeBrowser 2011
Supported species	3 (human, mouse, rat)	51 (new species like Drosophila melanogatser, Saccharomyces cerevisiae)
Number of annotations sources	19	54
Biological evidences considered	DAVIDknowledgebase (version 2005)	DAVIDknowledgebase (version 2007) and NEW ANNOTATIONS : – MicroRNA target site prediction (TargetScan, Pictar) – TFBS prediction (TFBSConserved, Cisred) – Protein-protein interaction – functional relationship (KEA) – Disease – Expression signatures (MsigDB, TBrowser's TS) – TBMC
Number of microarray platforms (i.e. GPL)	70	101
Number of microarray experiments (i.e. GSE)	1484	5568
Input for enrichment analysis	TS (more than 10 samples)	TS (more than 8 samples)
Generation of TS	bash, perl, C programms	optimisation DBF-MCL parameters (using RTools4TB) and an automatic pipeline (C, gawk, R)
Number of TS	18250	40138 (with 33941 ES corresponding to : Homo sapiens, Mus musculus and Rattus norvegicus)
% of annotated TS	84% annotated TS	87% annotated TS
Plugins	Heatmap, TBCommonGenes, TBMap	+ AnnotationOverview, TBConvertor, InteractomeBrowser
Request mode	boolean request by geneSymbol, probeID, platformID, experimentID, ES ID, ontologyID	new boolean request type : homologeneID, entrezID and request by list
Programmatic acces	NONE	Webservice : SOAP/WSDL and RTools4TB : R/Bioconductor package (DFB-MCL algorithm and webservice interface)

TABLE 4.1 – Bilan des avancées du projet entre sa publication en 2008 et maintenant

Il est ainsi possible d'utiliser des données de coexpression de gènes comme celles contenues dans notre base de données, afin d'améliorer la détection de bons candidats.

Autres approches de méta-analyses à partir des données présentes dans GEO

TBrowser n'est pas le seul projet ayant pour but d'étudier les coexpressions de gènes à partir de données de puces à ADN présentes dans les bases de données publiques comme GEO.

D'autres outils proposent des approches différentes (Table 4.2), mais toutes contrairement à TBrowser utilisent les informations sur les échantillons déposées dans GEO. Ils utilisent donc non pas les GSE comme TBrowser mais les *datasets* (GDS). Ils proposent d'obtenir à partir d'un gène donné des gènes similairement exprimés en nous renseignant sur le contexte expérimental.

GeneChaser permet de définir les différents contexte dans lesquels un gène donné est trouvé comme différentiellement exprimé, alors que MARQ renvoie une liste de gènes similairement différentiellement exprimés. D'autres proposent de construire des graphes de coexpressions (GEN-TREND). Ces outils proposent également parfois en entrée d'utiliser deux listes de gènes, correspondant à des gènes surexprimés et sousexprimés. L'objectif de ces outils est de nous renseigner sur les coexpressions de gènes, ils n'étudient pas les régulation autours de ces gènes en intégrant d'autres sources de données comme TBrowser.

Outil	Site web
TranscriptomeBrowser [Lopez *et al.*, 2008]	http://tagc.univ-mrs.fr/tbrowser
GENE CHAnge browSER (GeneChaser) [Chen *et al.*, 2008]	http://genechaser.stanford.edu/
MARQ [Vazquez *et al.*, 2010]	http://marq.dacya.ucm.es
Gene Expression data Mining Toward RElevant Network Discovery (GEM-TREND) [Feng *et al.*, 2009]	http://cgs.pharm.kyoto-u.ac.jp/services/network/index.php
COXPRESdb [Obayashi & Kinoshita, 2011]	http://coxpresdb.jp/
GOEGLE [Yu *et al.*, 2009]	http://omics.biosino.org:14000/kweb/workflow.jsp?id=00020
Genevestigator [Hruz *et al.*, 2008]	https://www.genevestigator.com/gv/biomed.jsp

TABLE 4.2 – Autres approches de méta-analyses de données de puces à ADN provenant de GEO, en gras l'outil que j'ai développé. Les cellules grisées correspondent aux outils non gratuits

Futurs développements et amélioration du projet TranscriptomeBrowser

Dans un premier temps, maintenant que la preuve de concept de notre algorithme DBF-MCL a été publié et que nous avons développé des plugins permettant l'analyse des informations disponibles dans notre base de données, il reste encore à publier la mise à jour.

Puis il est toujours possible d'améliorer encore la rapidité, la conception ou même encore les fonctionnalités de certains plugins : comme TBNeighborhood par la représentation graphique des coexpression par exemple. En effet, à l'heure actuel la matrice générée doit être analysé par d'autres outils.

L'étape suivante pourrait être l'intégration de nouvelles sources de données, avec pourquoi pas des données de puces à microARN, ou tout simplement de données de puces à ADN provenant de nouvelles plateformes non intégrée en 2009. Ceci nécessiterait une nouvelle mise à jour de la base, mais disposant de script générant automatiquement les données, cela ne devrait pas prendre beaucoup de temps. De plus, avec les bases de données comme lncRNAdb, de nouvelles informations concernant la régulation de l'expression des gènes pourra être obtenue.

La librairie R/Bioconductor RTools4TB nécessite quelques améliorations et mise à jour. Comme dit précédemment, l'intégration du nouveau service web SOAP/WSDL doit être finalisé. Ceci permettra une utilisation de la nouvelle base de données. De plus, d'autres fonctionnalités comme la générations de graphiques pour cytoscape ou pour une intégration dans InteractomeBrowser pourrait être intéressant. Enfin, la possibilité d'utiliser d'autres librairies R nous permettrait de faire des méta-analyses plus simplement et d'y intégrer des données analysées sous R et provenant de techniques à très haut débit comme le RNA-seq ou le ChIP-seq. On s'oriente ainsi vers une intégration de plusieurs types d'expériences comme le propose déjà l'outil Genomics Portals ([Shinde *et al.*, 2010]).

Chapitre 5

Étude de la régulation transcriptionelle par HTS

Sommaire

La complexité des mécanismes de régulation transcriptionnelle commence seulement à être élucidée. Dans de nombreux tissus ou types cellulaires, des facteurs de transcription essentiels au fonctionnement normal ou pathologique ont été identifiés, mais seules quelques unes de leurs cibles directes sont connues. De nos jours, de nombreuses techniques permettent d'étudier la régulation de l'expression des gènes à grande échelle et à très haut débit (voir Chapitre 1). Ainsi, le ChIP-seq ([Johnson *et al.*, 2007], [Mardis, 2007], [Elnitski *et al.*, 2006], [Massie & Mills, 2008]), qui associe immunoprécipitation de la chromatine (« *Chromatin*

ImmunoPrecipitation » ou ChIP) ([Gilmour & Lis, 1985]) et séquençage à très haut débit des fragments d'ADN immuno-précipités, permet l'identification précise, *in vivo*, des sites de fixation des protéines dans le génome. Ces protéines se liant à l'ADN comprennent notamment les ARN polymérases, les facteurs de transcription et les histones.

5.1 Principe de l'immunoprécipitation de la chromatine associée au séquençage à très haut débit (ChIP-seq)

5.1.1 Généralités

L'arrivée récente du HTS a révolutionné l'étude à grande échelle des mécanismes de régulation de l'expression des gènes. Associée au ChIP, cette technique constitue un outil extrêmement efficace pour (1) la détermination des sites de fixation des facteurs de transcription (de manière directe ou indirecte dans le cas des co-facteurs), (2) la localisation des modifications des domaines N-terminaux d'histones, et (3) l'étude de la fixation de diverses protéines sur l'ADN (ARN polymérase, variants d'histone).

Anciennement hybridés sur des puces à ADN pangénomiques (« *tiling arrays* ») par la technique de ChIP-on-chip ([Blat & Kleckner, 1999], [Ren *et al.*, 2000], [Robyr *et al.*, 2002]), les fragments d'ADN immunoprécipités sont désormais séquencés pour permettre d'identifier des sites de fixation des facteurs de transcription de façon plus précise (Figure 5.1 et Table 5.1). Diverses techniques utilisant un séquençage par la méthode de Sanger ont ainsi vu le jour : *ChIP-serial analysis of chromatin occupancy* (SACO, [Impey *et al.*, 2004]), *ChIP-serial analysis of binding elements* (SABE, [Chen & Sadowski, 2005]), *ChIP-sequence tag analysis of genomic enrichment* (STAGE, [Bhinge *et al.*, 2007]), *Genome-wide mapping technique* (GMAT, [Roh *et al.*, 2004]).

Ces approches ont récemment été supplantées par le ChIP-seq, une technique associant immunoprécipitation de la chromatine et séquençage à très haut débit des fragments de sonication immunoprécipités ([Barski *et al.*, 2007], [Johnson *et al.*, 2007]). Contrairement aux techniques antérieures, le ChIP-seq permet de déterminer le site de fixation d'une protéine avec une précision de quelques dizaines de bases seulement, pour peu que la couverture (nombre de fragments couvrant la zone d'intérêt) soit suffisante ([Ho *et al.*, 2011]) (Figure 5.1 et Table 5.1). De plus, l'utilisation du mode de séquençage « paired-end » a permis d'accroître encore la spécificité et la précision des résultats (ChIP-PET) ([Zeller *et al.*, 2006]).

Enfin, on peut mentionner l'existence d'une nouvelle technique à très haut débit appelée ChIA-PET (Chromatin Interaction Analysis using Paired End Tag sequencing). Combinant ChIP-PET et 3C-seq (voir partie 1.4.2), elle a été récemment utilisée pour l'étude des enhancers, régions régulatrices éloignées des régions promotrices et géniques ([Fullwood *et al.*, 2009]).

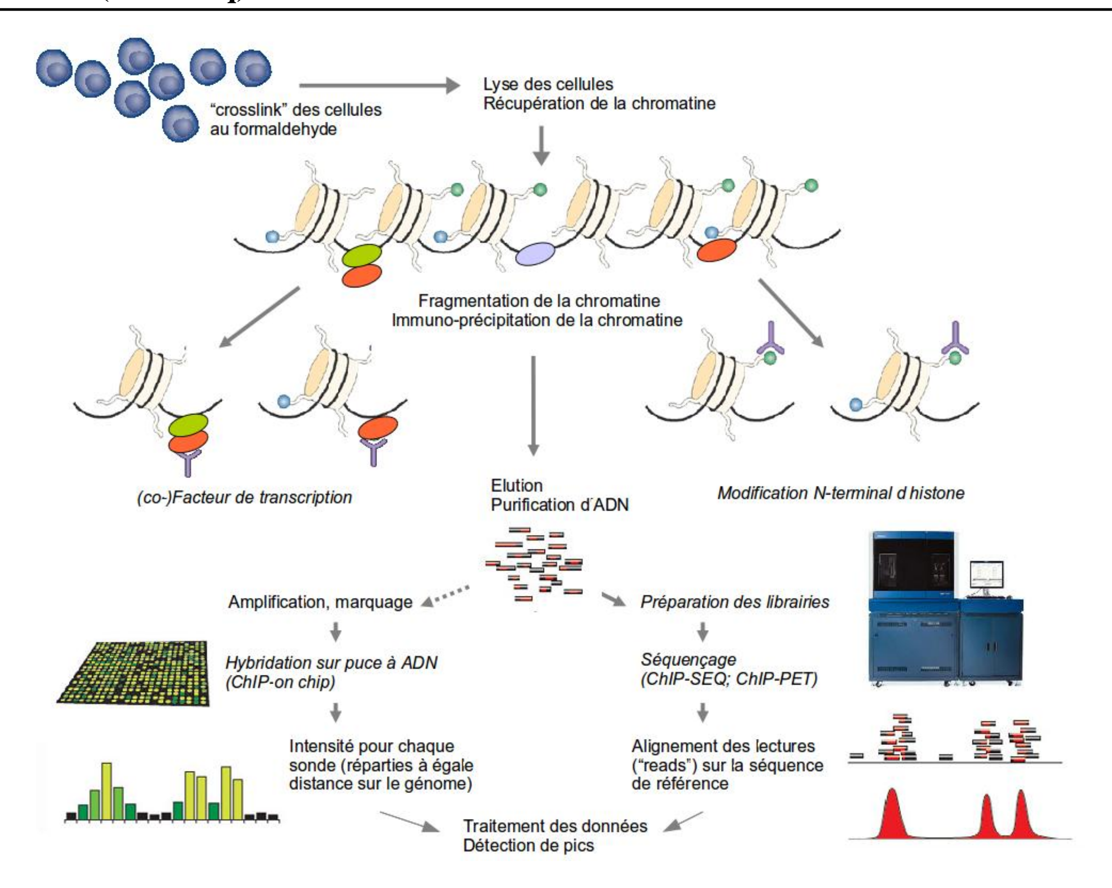


FIGURE 5.1 – ChIP-seq vs. ChIP-on-chip : processus général

Technologie	ChIP-on-chip (puces à ADN)	ChIP-seq (séquençage à très haut débit)
Révélation	Hybridation d'ADNc	Séquençage d'ADNg
Résolution	6.5 millions de sondes par puce	> 700 millions de séquences obtenues (*reads*) par run
Couverture du génome	Limitée par le nombre de sondes présentes sur la puce	Illimitée
Risque de cross-hybridation	Oui, entre les séquences très similaires	Aucun
Multiplexage	Non	Oui
Taille des fragments	600 pb	150-300 pb
Nombre de cellules	5 x 10^7 à 1 x 10^8	10 x 10^6 à 1 x 10^7 (* > 1 x 10^5)

TABLE 5.1 – comparaison des techniques de ChIP-on-chip et de ChIP-seq. Le « * » correspond à l'utilisation du kit MAGnify™

5.1.2 Principe biologique

La première étape du ChIP, appelée « crosslink », permet de fixer les protéines à l'ADN de manière covalente, afin de pouvoir étudier leur interaction sur tout le génome. Une fois produites en quantité suffisante, les cellules subissent un traitement au formaldéhyde qui a pour objectif de créer des liaisons covalentes entre les résidus de lysine des protéines et les cytosines de l'ADN, tout en préservant l'intégrité structurelle des cellules. Cette réaction, stoppée par l'ajout de glycine, est réversible. Elle est l'une des étapes primordiales du ChIP.

Les cellules traitées au formaldéhyde subissent des lyses successives afin d'extraire la chromatine. Celle-ci est ensuite fragmentée, soit par sonication, soit par digestion enzymatique, afin d'obtenir des fragments d'ADN « crosslinkés » aux protéines, de la taille souhaitée. Pour l'étude d'un facteur de transcription, plus les fragments seront petits (longueur comprise généralement entre 150 et 300 nucléotides), plus la définition des positions des sites de fixation sera précise. Pour les marques d'histones, la taille de fragment est de 146 nucléotides, soit la taille d'un nucléosome ; elle est obtenue par digestion à la Micrococcal Nuclease (MNase) permettant la coupure entre les nucléosomes.

Les fragments de chromatine sont ensuite immuno-précipités à l'aide d'un anticorps spécifique du facteur de transcription étudié ou d'une modification d'histone. Ces anticorps sont couplés à des billes magnétiques à l'aide de protéines G ou A (Dynabeads protein G, Life Technologies) choisies en fonction du sérotype de l'anticorps utilisé. Une petite quantité de chromatine non immunoprécipitée est conservée : elle est appelée « *input* ». Elle permet de mesurer la fixation non spécifique et donc de corriger le bruit de fond des données de séquençage. Précédemment, l'input était obtenu par immunoprécipitation de la chromatine à l'aide d'immunoglobuline de même sérotype que l'anticorps utilisé pour le ChIP.

Les fragments de chromatine immunoprécipités et l'input sont traités à la protéase K et à la RNAse afin de supprimer l'effet du crosslink. Enfin, l'ADN est extrait et purifié pour permettre la préparation des librairies de séquençage.

5.1.3 Biais et bruit de fond

Pour limiter les biais et réduire le bruit de fond, il est important de prendre en compte quelques contraintes techniques.
En effet, les étapes de crosslink et de sonication nécessitent une mise au point technique dont dépend la qualité des résultats. Il faut, en fonction des cellules utilisées, veiller à adapter divers paramètres tels que la durée de crosslink et de sonication, le pourcentage de formaldéhyde, l'intensité de sonication (selon le modèle du sonicateur), le volume final, la quantité de cellules. Autant de paramètres qui détermineront la qualité du ChIP et l'obtention de fragments de la taille souhaitée. La taille des fragments d'ADN pour le ChIP-seq est en général comprise entre 100 et 300 paires de bases en fonction des paramètres de sonication (temps, intensité). Elle est vérifiée par migration sur gel SDS-page ou à l'aide du bioanalyzer Agilent.

L'abondance des protéines fixées ou des histones modifiées, ainsi que la qualité de l'anticorps sont des critères qui doivent être pris en compte afin de déterminer le nombre optimal de cellules qui sont nécessaires pour l'expérience. Le rapport signal sur bruit étant directement corrélé à la quantité d'ADN, l'utilisation d'un nombre excessif de cellules tend à augmenter le bruit de fond ([Kidder *et al.*, 2011]). Ainsi, pour une expérience de ChIP-seq, le nombre de cellules utilisées est généralement compris entre $1x10^6$ et $10x10^6$, ce qui équivaut à 10-100 ng d'ADN immunoprécipité. De petites quantités de cellules sont généralement suffisantes pour l'analyse de protéines abondamment fixées telles que l'ARN polymérase II et pour la localisation de certaines modifications d'histones comme la tri-méthylation de la lysine 4 de l'histone 3 (notée « H3K4me3 »), alors que de très grosses quantités seront nécessaires pour des protéines moins abondantes ou des modifications d'histones plus diffuses. Il existe des kits permettant de travailler sur de faibles quantités de cellules, comme le kit MAGnify™de Life Technologies avec lequel on peut utiliser de $1x10^4$ à $1x10^5$ cellules.

L'un des points les plus cruciaux dans la réalisation d'un ChIP est bien entendu le choix de l'anticorps. Il devra avoir une forte affinité et une spécificité élevée envers l'épitope de la protéine étudiée pour permettre un enrichissement spécifique et donner au séquençage un rapport signal sur bruit statistiquement significatif. On choisira préférentiellement un anticorps monoclonal (se liant à un seul épitope) afin d'éviter toute réaction d'hybridation parasite.

Dans le cas où les anticorps dirigés contre un facteur de transcription donné ne sont pas utilisables, car ils ne donnent pas de bons résultats en ChIP, il est parfois possible de faire exprimer par des cellules les protéines d'intérêt marquées à l'aide de « tags » tels que Myc, HA (hemagglutinine) ou bien encore la biotine. L'anticorps utilisé pour réaliser le ChIP sera alors un anticorps spécifique du tag et non plus de la protéine. De plus, ces tags possèdent une meilleure spécificité (surtout pour la biotine en interaction avec la streptavidine). L'inconvénient reste que cette modification des protéines peut altérer le système biologique étudié.

Avec ces techniques, la présence d'artefacts est un sujet de préoccupation. Il est donc essentiel de réaliser des contrôles appropriés. Les contrôles internes de l'expérience peuvent être divers : input, anticorps non spécifique (Immunoglobuline de même sérotype), ChIP sur un autre tissu, réplicats techniques ou biologiques utilisant d'autres anticorps spécifiques de la protéine. Cependant, le coût élevé du séquençage d'un échantillon freine leur utilisation.

Il existe un dernier contrôle capital avant de passer au séquençage : vérifier que le ChIP a produit un enrichissement suffisant (20% d'enrichissement minimum). Cette vérification peut être faite à condition de connaître un gène cible (ou site de fixation) du facteur de transcription. En comparant les résultats de qPCR de l'échantillon immunoprécipité et du contrôle, on peut établir l'enrichissement en séquences d'intérêt.

5.1.4 Avantages et inconvénients

Le ChIP-on-chip et le ChIP-seq sont actuellement deux technologies utilisées pour l'étude de la fixation des protéines sur l'ADN. Toutefois, le ChIP-seq supplante le ChIP-on-chip en raison d'avantages notables ([Massie & Mills, 2008]) :

- possibilité d'analyse du génome entier, il ne dépend pas des sondes présentes sur la puce à ADN. En effet, les puces à ADN disponibles possèdent un nombre limité de sites qui ne représentent qu'une fraction du génome total ;
- meilleure sensibilité (toutes les séquences présentes sont séquencées) et reproductibilité ;
- possibilité de multiplexage (utilisation de code-barres) ;
- possibilité de séquencer en mode « *paired-end* » pour améliorer la qualité de l'alignement des fragments ;
- suppression des biais techniques liés aux puces à ADN comme la saturation du signal, les problèmes de détection de spots, de lavage partiel de la puce ;
- besoin d'une quantité initiale d'ADN plus faible, ce qui est pratique pour des échantillons précieux, avec au minimum 5μg pour le chip-on chip contre 5ng pour le ChIP-seq ;
- moins de bruit de fond avec une gamme dynamique et un rapport signal sur bruit plus précis (pas de bruit de fond des spots) et l'absence d'hybridation croisée entre les sondes ([Johnson *et al.*, 2007], [Mardis, 2007], [Massie & Mills, 2008]) ;
- meilleure résolution spatiale des pics ou profils : un site de liaison d'un facteur de transcription peut être identifié précisément (10-30 pb centré sur le pic) ([Kharchenko *et al.*, 2008]).

Cependant le ChIP-seq possède également quelques inconvénients :

- son coût. Néanmoins, cet inconvénient est atténué par le fait que ce mode de séquençage donne beaucoup plus d'informations que le ChIP-on-chip. En outre, cette différence de coût s'estompe progressivement avec les avancées technologiques ;
- temps de préparation des librairies plus long que celui de la préparation des échantillons pour l'hybridation sur puce à ADN avec des protocoles plus complexes ;
- l'amplification PCR (Figure 1.7) de ces HTS conduit parfois à des biais d'amplification, certains *reads* sont ainsi sur-amplifiés tandis que d'autres sont sous-amplifiés ([Mutter & Boynton, 1995]) ;
- erreurs de séquençage (région composée d'un faible pourcentage en GC ([Siddiqui *et al.*, 2006]) et d'alignement possibles (séquences répétées et régions télomériques) ([Dohm *et al.*, 2008]) ;
- temps d'analyse plus long et besoin de plus de moyens informatiques pour le stockage et le traitement d'un volume important de données (en Gb pour le ChIP-seq plutôt qu'en Mb pour le ChIP-on-chip) ;
- analyse de données plus complexes (modèles statistiques, normalisation des données).

5.1.5 Le modèle théorique de distribution des séquences

Après alignement des séquences immunoprécipitées sur le génome (ou une séquence de référence), deux types de distributions peuvent être étudiées en fonction du type d'expérience :

- la localisation de sites de liaison d'un facteur de transcription ;
- l'étude de la distribution des marques épigénétiques le long des gènes du génome.

Les sites de liaison des facteurs de transcription sont détectés par l'observation d'un enrichissement localisé en séquences immunoprécipitées, formant un pic à l'endroit précis où le facteur est lié à l'ADN ([Wilbanks & Facciotti, 2010], [Pepke *et al.*, 2009]) (Figure 5.2). Les fragments de chromatine immunoprécipités à l'aide de l'anticorps spécifique de cette protéine sont de taille variable, comprise entre 150 et 300 nucléotides. Pour chacun de ces fragments, le séquençage permet l'acquisition de la séquence des 50 premiers nucléotides (si on utilise un séquenceur de type SOLiD en mode fragment). Après alignement sur la référence, ces 50 nucléotides doivent mettre en évidence un déséquilibre de brin avec, au centre, le site de liaison du facteur de transcription (Figure 5.2). Le déséquilibre de brin est schématiquement représenté par la présence d'un pic sur chaque brin (+ et -) du génome. Ces pics sont décalés d'une distance d correspondant à la taille de sonication.

Le facteur de transcription étudié peut être soit en interaction avec la chromatine ou soit directement lié à l'ADN au niveau d'un site séquence-spécifique (appelé motif).

Les marques épigénétiques (modifications d'histones, variants d'histones, ou bien encore ARN polymérases), sont étudiées en analysant leur distribution au niveau des gènes, par recherche de profils différentiels autour du site d'initiation de la transcription (TSS), ou tout au long du gène, dans diverses conditions expérimentales ou pour différents types cellulaires ([Kidder *et al.*, 2011], [Barski *et al.*, 2007]) (Figure 5.3). Cette différence peut également être étudiée de manière globale (recherche de différentiel de couverture) sur l'ensemble du génome, dans le but d'identifier des enhancers. En effet, il a récemment été montré que les sites de liaison intergéniques de l'ARN polymérase II étaient localisés de façon préférentielle à proximité des enhancers ([De Santa *et al.*, 2010]).

5.2 L'informatique du HTS

Le développement rapide des technologies HTS implique la mise en place d'outils et méthodes performants pour l'analyse des données et leur mise à jour permanente. Chaque séquenceur possède une architecture système propre ainsi que des suites logicielles propriétaires spécifiques avec des formats de données non standards.

Le matériel nécessaire au fonctionnement des séquenceurs HTS et à l'analyse des données générées est donc très important. En effet, il ne suffit pas de disposer d'un ordinateur de bureau comme pour les technologies de puces à ADN (donnes un ordre de grandeur des fichiers de données) ; les données de séquençage à très haut débit génèrent un flux de données qui se mesure en gigabytes (Gb) voire en terabytes (Tb).

5.2.1 Organisation matérielle et logicielle

Afin de pouvoir aligner les séquençages et analyser les résultats générés par une expérience, il est indispensable de posséder une ferme de calcul puissante (on utilise communément le terme anglais « *cluster* ») et des unités de stockage dédiées qui permettront de stocker les Tb

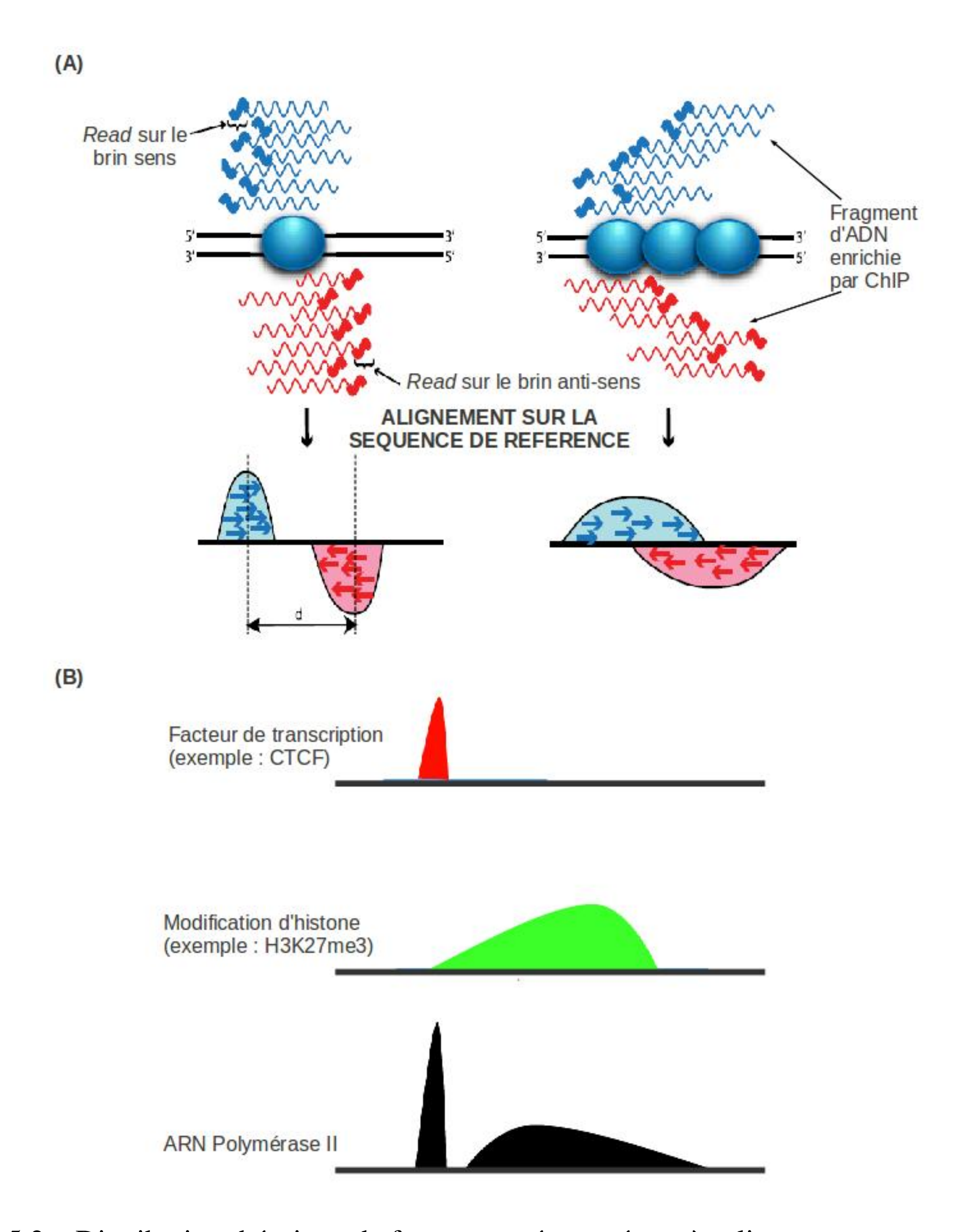


FIGURE 5.2 – Distribution théorique de fragments séquencés après alignement sur une séquence de référence avec en (A) la définition d'un pic, où d correspond à la taille de sonication et en (B) les différents profils de pics (Adapté de [Wilbanks & Facciotti, 2010] et [Kidder *et al.*, 2011])

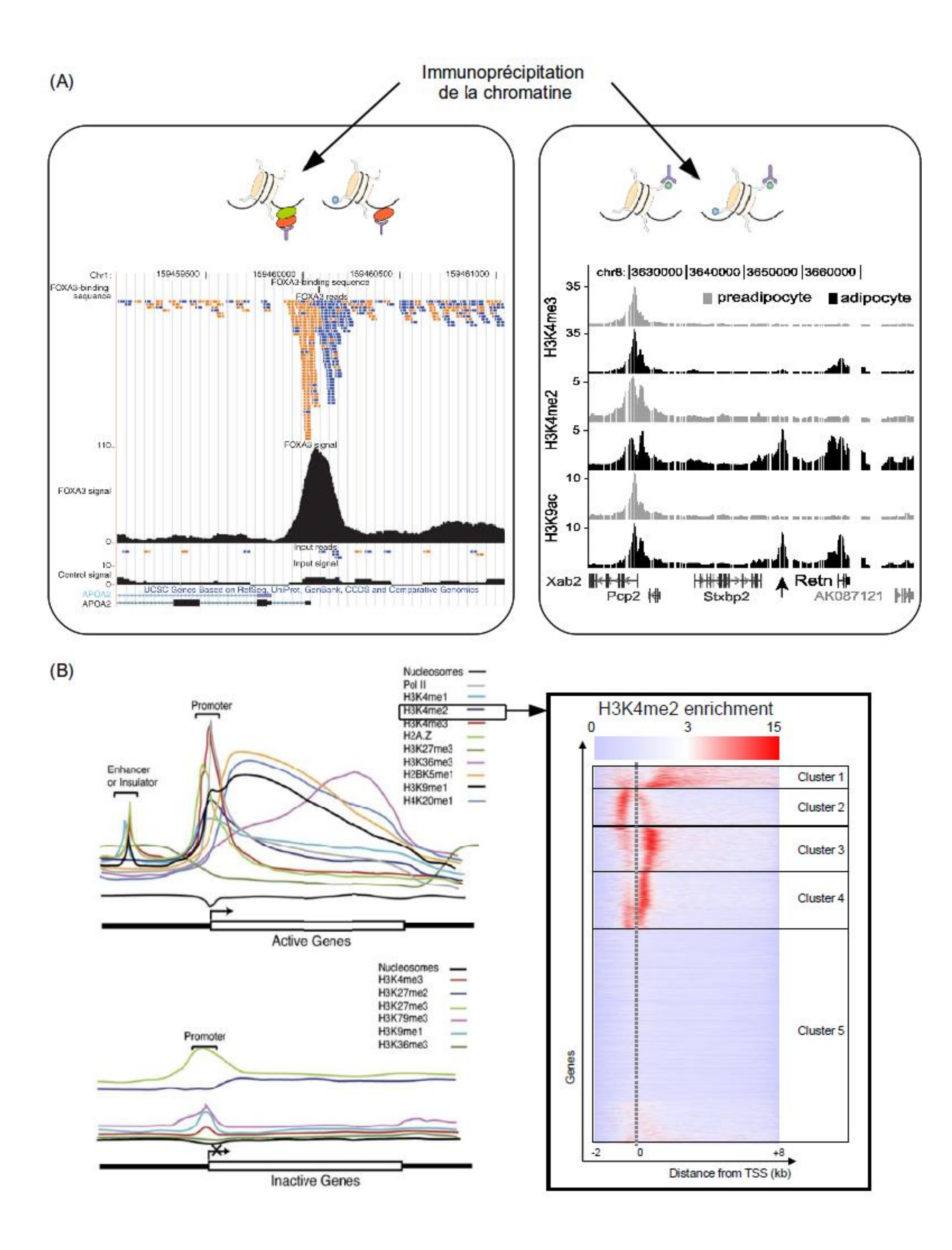


FIGURE 5.3 – Visualisation de profils de pics avec en (A) ceux obtenus pour un facteur de transcription ou pour les marques de méthylation et en (B) la différence de profils entre les différentes modifications d'histone ([Barski *et al.*, 2007], [Tomaru *et al.*, 2009] et [Pekowska *et al.*, 2010])

de données et de les réutiliser facilement si nécessaire. Le SOLiD™est livré avec un cluster de calcul (« *cluster online* ») permettant : (1) l'acquisition des images de chacune des 10 ligations pour les 5 cycles d'amorces, (2) le stockage des images et (3) pour chaque bille, la détermination de la séquence en code couleur puis sa conversion en nucléotides. Il faut donc un matériel performant pour analyser ces images, et une grosse capacité de calcul pour gérer les millions de billes déposées sur les lames ; la nécessité de performance ne provient pas tellement de la complexité des calculs à réaliser, mais souvent du nombre de fois qu'il faut les réaliser.

Afin de concilier simultanément l'acquisition et le traitement des données de séquençage, un second cluster a été installé au TAGC (« *cluster offline* »). Celui-ci possède la même organisation matérielle, ainsi que le même environnement logiciel que le *cluster online* (Figure 5.4). Cependant, il offre une puissance de calcul supérieure grâce à des processeurs plus récents. Ces 2 clusters sont composés d'un noeud maître (« head node ») et de quatre noeuds esclaves (« node »). On appelle ce type de ferme de calcul, un « cluster Beowulf ». Leur système d'exploitation, suivant les pré requis de Life Technologies, est CentOS (« *Community ENTerprise Operating System* »), une distribution GNU/Linux gratuite, dérivée de Red Hat, principalement destinée aux serveurs. Ce système n'est physiquement installé que sur le noeud maître des deux serveurs, les quatre noeuds esclaves chargeant leur système en mémoire au démarrage grâce au logiciel Scyld. Enfin, le logiciel Torque (« *Terascale Open-source Resource and QUEue Manager* »), version gratuite du logiciel PBS (« *Portable Batch System* »), associé à un séquenceur de tâche, permet la gestion par le maître de la répartition des tâches sur les noeuds du cluster, et l'envoi de commandes pour la soumission des jobs et leur monitoring. Au total, ces deux clusters possèdent 40 coeurs de calculs ; par comparaison, un ordinateur de bureau comporte généralement 2 coeurs (« dual core »).

Pour la conservation des données à moyen et long terme, le laboratoire dispose de 4 unités de stockage en réseau, également appelé stockage en réseau NAS (de l'anglais « *Network Attached Storage* »). Ainsi, deux Dell MD1000 de 13 Tb chacun, directement connectés en SAS aux clusters de calculs, sont utilisés pour l'acquisition et l'analyse des données. Une baie NetApp de 8 To est dédiée au stockage des résultats et une unité Netgear de 20 To contient les résultats bruts de séquençage. Enfin, le stockage des données brutes à très long terme est assuré par une sauvegarde sur bande LTO4. Nous disposons donc d'un total de 50 Tb d'espace de stockage. La production du séquenceur est telle qu'en un an, la totalité des 50Tb de stockage a été utilisée. En effet, une expérience (ou « *run* ») et son analyse génèrent 1 à 2 Tb de données en fonction du mode de séquençage et du type d'analyse.

Enfin, différentes suites logicielles permettent de contrôler le séquençage et l'analyse des données du SOLiD™(Figure 5.5). Elles correspondent respectivement à ICS (« *Instrument Controler Software* ») / SETS (« *SOLiD Experimental Tracking Software* ») et Bioscope / Corona Lite. Des programmes et scripts développés au laboratoire ou des open-sources sont également nécessaires pour permettre l'analyse de l'immense flot de données. Enfin, les navigateurs de génomes, communément appelés *Genome Browser*, permettent l'affichage de l'alignement des *reads* le long d'un génome, ainsi que d'autres annotations telles que les gènes, les transcrits, les séquences répétées.

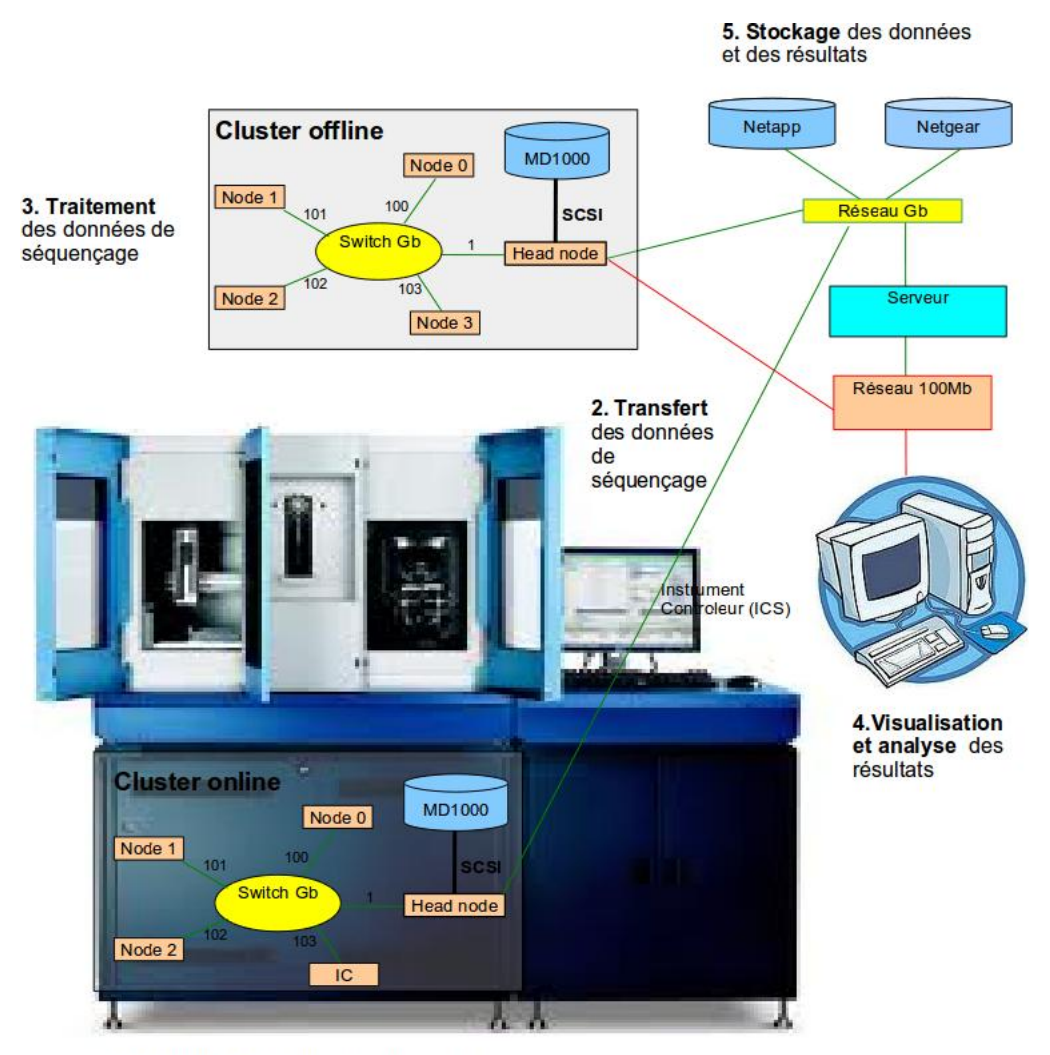


FIGURE 5.4 – Schéma de l'organisation du matériel utilisé sur la plateforme IBiSA TGML du TAGC pour l'acquisition et l'analyse des données de séquençage à très haut débit par la technologie SOLiD^{TM}

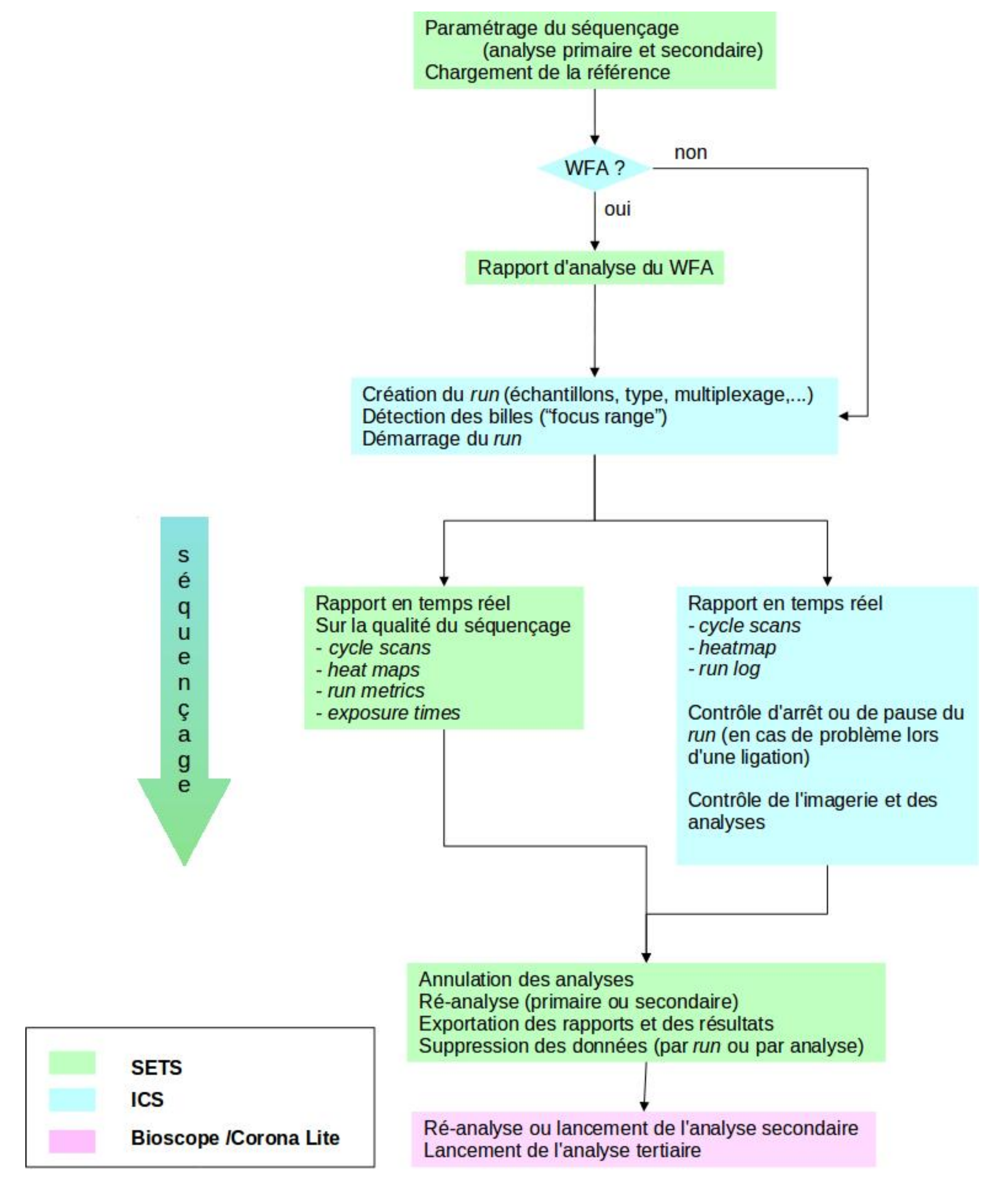


FIGURE 5.5 – Déroulement de la préparation du séquençage et de l'analyse à l'aide des différents logiciels (Adapté du manuel d'utilisation d'Applied Biosystems « SOLiD Experimental Tracking Software (SETS) v4.0.1 »)

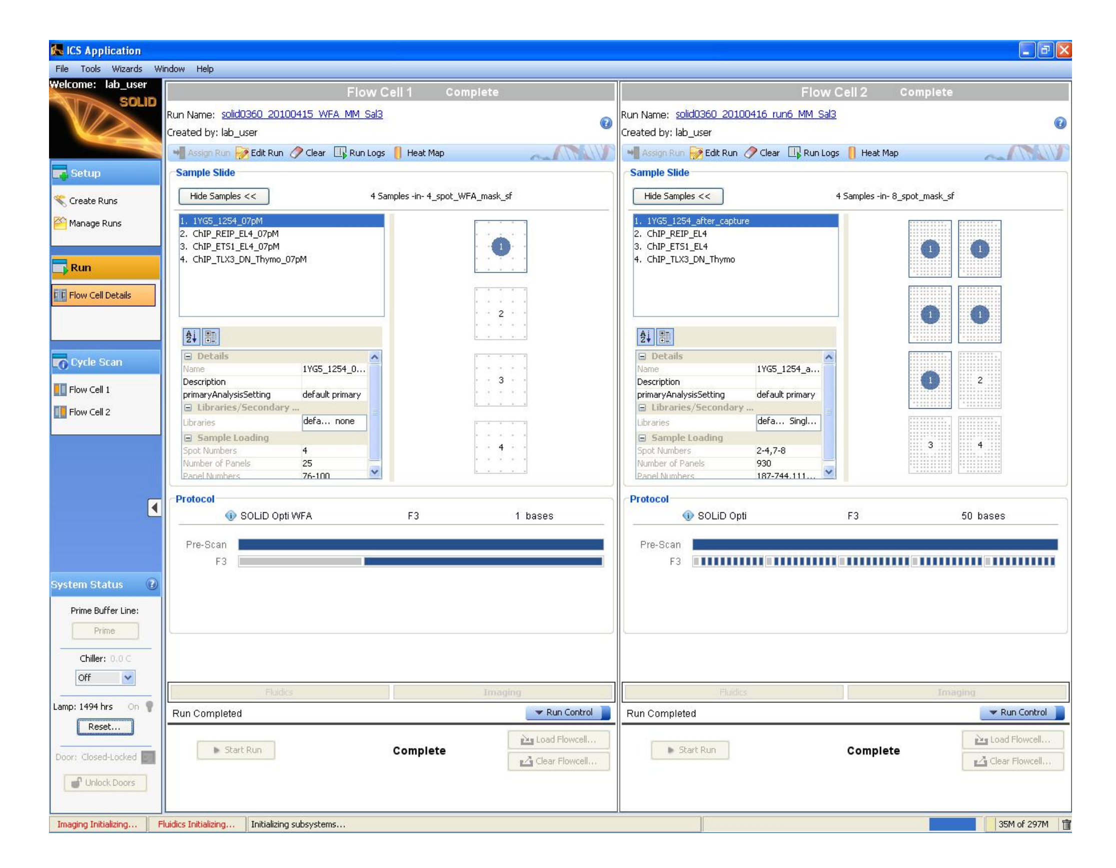


FIGURE 5.6 – Aperçu de l'interface graphique du logiciel ICS pilotant le run

5.2.2 Interfaces utilisateurs pour le lancement et la gestion du séquençage

Avant de lancer une session de séquençage, il est nécessaire de paramétrer le séquenceur et les analyses à l'aide de ICS et SETS. ICS est le logiciel de contrôle instrumental du SOLiD^{TM}, tandis que SETS est le logiciel de gestion des séquençages sur le cluster online (Figure 5.6). C'est une application web qui permet la visualisation des données en temps réel et la lecture des rapports d'analyse une fois le run terminé.

5.2.3 Pipeline de traitement de données : Bioscope

Bioscope est une suite logicielle développée par Applied Biosystems et livrée en standard avec les séquenceurs SOLiD^{TM}. Elle est utilisée pour réaliser l'analyse secondaire et certaines analyses tertiaires telles que la recherche de SNP, de petits et larges *indels*, d'inversions, de CNV (Copy Number Variations) ou encore le calcul de l'abondance des transcrits après séquençage d'un exome (*whole exome*). Son fonctionnement, en ligne de commande, se fait via des fichiers de configuration *.ini contenant tous les paramètres des analyses que l'on souhaite faire (Figure 5.8). Une base de données Postgres est utilisée pour la gestion des tâches, ainsi que le système de messagerie Java ActiveMQ, ce qui permet de lancer plusieurs analyses en même temps. Bioscope possède également une interface graphique écrite en Java qui s'exécute dans le conteneur Tomcat (Figure 5.7). Cette interface, très pratique, ne sert en fait qu'à produire les fichiers ini pour Bioscope, et à lancer ce dernier. Cette interface Bioscope

autorise également l'ajout de fonctionnalités par le biais de modules (ou « *plugins* »), comme par exemple ceux de Corona Lite.

Corona Lite est une suite logicielle open source en ligne de commandes maintenue par Applied Biosystems. Elle permet l'alignement (« *mapping* ») des *reads* par comparaison à une référence, et la détection des petits *« indels »* et des SNP. Elle est développée en Perl, Python et Java.

5.3 Analyse de données de ChIP-seq

Comme pour les puces à ADN il y a une quinzaine d'année, l'apparition du ChIP-seq a nécessité le développement de nombreux outils et méthodes spécifiques d'analyse. En effet, la masse de données à analyser pour chaque expérience requiert des outils de plus en plus performants. Ces outils mettent en oeuvre divers principes et méthodologies et sont parfois liés à une technologie de séquençage ou même à un format de fichier particulier.

Les principales étapes de l'analyse de données de ChIP-seq seront présentées ici, en mettant l'accent sur les données issues de la technologie SOLiD™, dont l'analyse a fait partie de mes travaux de thèse.

Quelle que soit la technique HTS utilisée, l'analyse des données est composée de trois étapes successives :

- l'**analyse primaire** correspond à l'acquisition des images de séquençage (4 par ligation, car 4 couleurs), à partir desquelles deux fichiers sont produits, l'un contenant la séquence de chaque bille en code couleur (« *color call* ») et l'autre la qualité (« *quality metric* ») de chaque ligation par bille. Cette étape est la seule à être exclusivement réalisée par le cluster de calcul associé physiquement au séquenceur (« *cluster online* ») ;
- l'**analyse secondaire** est l'alignement de ces séquences en code couleur sur le génome ou la séquence de référence. Elle dépend de la technologie utilisée (en effet, les données SOLiD™ne sont pas au même format que celles des autres technologies) mais également du mode de séquençage (fragment, paired-end). Il est parfois indispensable de transformer le format des données pour pouvoir utiliser l'outil choisi ;
- l'**analyse tertiaire** dépend de la technique HTS et correspond à l'analyse spécifique des données, comme la recherche de pics pour le ChIP-seq. Elle dépend de divers facteurs tels que le mode de séquençage « *fragment* » ou « *paired-end* », la prise en compte de l'input (ou d'un autre contrôle), ou bien encore du type d'expérience (facteur de transcription ou modification d'histones).

Une fois ces analyses terminées, les alignements de séquences, les positions des pics et la distribution des *reads* sur le génome peuvent être visualisés et interprétés. En effet, à partir de cibles potentielles de facteurs de transcription, il est possible (1) de vérifier la présence de motifs de fixation du facteur de transcription sous les pics, (2) de définir un motif consensus dans des séquences extraites à partir des séquences sous les pics, (3) d'étudier les fonctions des gènes cibles et (4) de localiser des pics au sein du gène (intron/exon, UTR, région intergé-

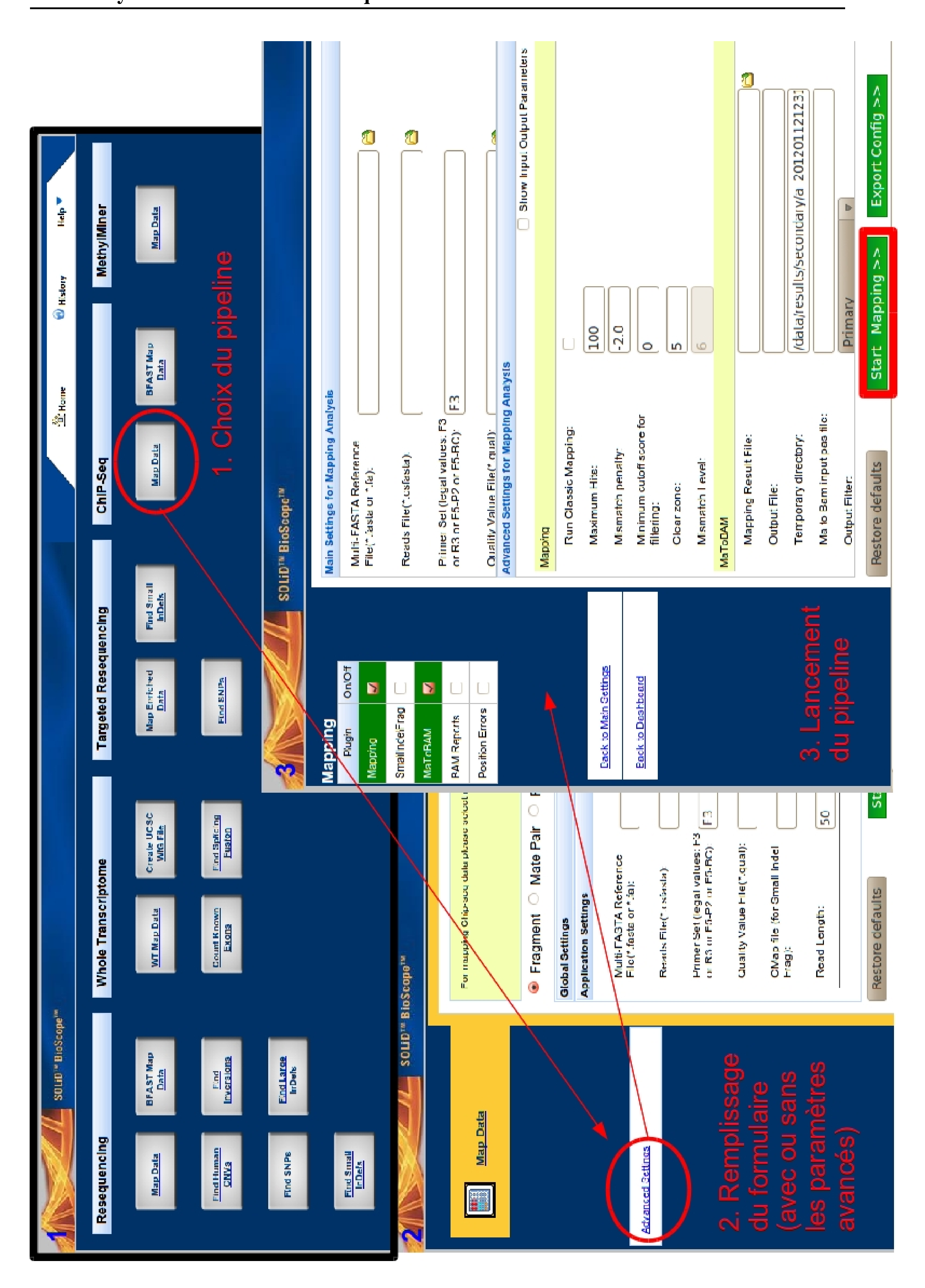


FIGURE 5.7 – Interface graphique Tomcat de Bioscope installée sur le cluster offline avec les différents pipelines disponibles et un exemple de lancement de pipeline

```
##  global settings for the pipeline run
import ../../globals/global.ini
reference = ${reference.dir}/chr20.fasta
run.name = myRun
sample.name = chr20
primer.set = F3
read.length = 50
output.dir = ${base.dir}/../outputs

##  qv filtering pipeline
classify.run = 1
read.dir = ${base.dir}/../../human_var/secondary/JOAN/mappingF3
read.file.prefix = ${run.name}_${sample.name}_${primer.set}
mapping.tagfiles.dir = ${output.dir}/qvfiltered
filtering.qv.filtered.dir = ${output.dir}/qvfiltered
filtering.qv.failed.dir = ${output.dir}/qvfail

##  mapping pipeline
mapping.run = 1
mapping.tagfiles.dir =${base.dir}/../../human_var/secondary/JOAN/mappingF3
mapping.output.dir = ${output.dir}/s_mappingF3
```

```
##  global parameters
base.dir=./
output.dir = ${base.dir}/outputs
temp.dir = ${base.dir}/temp
intermediate.dir = ${base.dir}/intermediate
log.dir = ${base.dir}/log
reads.result.dir.1 = ${base.dir}
reads.result.dir.2 = ${base.dir}
reference.dir =
/data/results/bioscope_examples/examples/references
scratch.dir=/scratch/solid
```

FIGURE 5.8 – Fichier de paramétrage (.ini) de bioscope pour une utilisation en ligne de commandes

nique). Ces résultats peuvent également être croisés avec d'autres types de données, tels que des données de transcriptome (puces à ADN, RNA-seq), de méthylation (Me-DIP), d'ouverture de la chromatine (FAIRE-seq), afin de construire des réseaux de régulation contextualisés.

Enfin, comme pour les puces à ADN, il faut passer par une étape de validation expérimentale par ChIP-qPCR de certaines positions cibles. Elle est généralement réalisée sur le ChIP de départ car seule une partie de l'ADN immunoprécipité est utilisée pour le séquençage. Mais ces validations peuvent également être réalisées sur un ChIP indépendant.

5.3.1 Données brutes et qualité de séquençage

Les séquences, qu'elles soient protéiques ou nucléiques, sont la plupart du temps stockées dans des fichiers « fasta ». Ces derniers, qui peuvent contenir une ou plusieurs séquences, sont des fichiers textes structurés de la façon suivante : chaque séquence débute par une ligne, dont le format est libre, commençant par un caractère « > », et contenant diverses informations telles qu'un identifiant de séquence ou un nom de gène. Les lignes suivantes, d'égale longueur, contiennent la séquence en nucléotides, acides aminés ou code couleur, dans le cas des fragments de séquençage produits par le séquenceur SOLiD^{TM}.

Dans le cas du séquençage à très haut débit, des fichiers de séquences un peu plus complexes ont été utilisés et sont maintenant devenus des standards. Outre les séquences, ces fichiers contiennent des valeurs de qualité pour chaque base séquencée. Les principaux formats de fichiers de séquences brutes (non alignées) disponibles à l'heure actuelle sont :
- le « csfasta » et le « _QV.qual », pour le SOLiD^{TM},
- le « sff » (« Standard Flowgram Format »), pour le 454 de Roche,
- le « fastq », pour l'Ion Torrent, l'Illumina et les autres technologies de séquençage à très

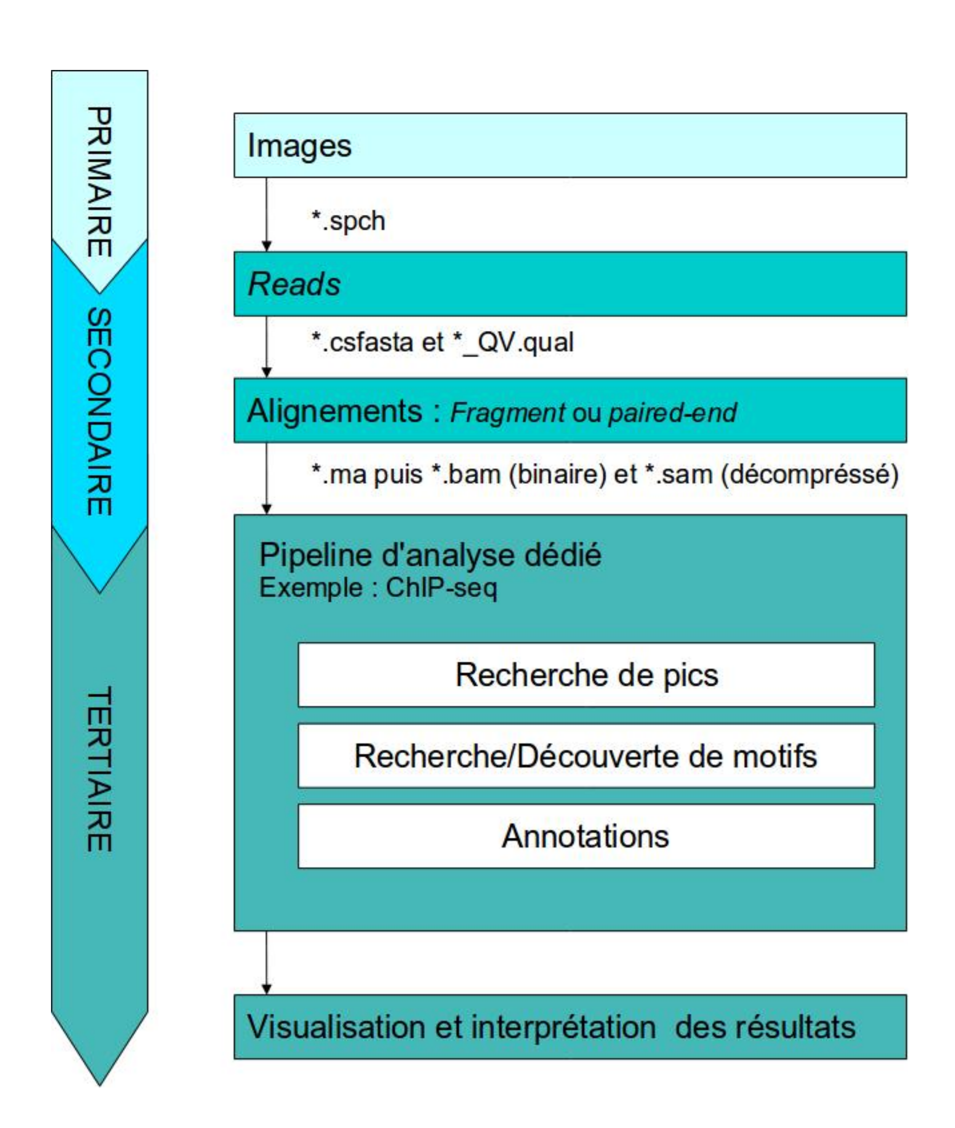


FIGURE 5.9 – Pipeline d'analyse des données de ChIP-seq

haut débit.

Ainsi, pour le SOLiD™, si le séquençage s'est bien déroulé, les images réalisées au cours de chaque ligation sont transformées en fichiers « spch », dans un format conçu pour des données volumineuses et complexes, appelé HDF5 (http://www.hdfgroup.org/HDF5/whatishdf5.html). Ces nombreux fichiers sont ensuite utilisés afin de créer, pour chaque échantillon, deux fichiers : un de séquences (.csfasta) et un contenant les scores de qualité pour chaque dibase (_QV.qual) (Figure 5.10). Il est à noter que le SOLiD™génère 2 fichiers au lieu d'un seul dans le cas des autres technologies. Ils contiennent une en-tête commentée en « # » pour mettre les lignes de commandes avec les paramètres ayant permis de les générer. Ces fichiers incluent toutes les séquences, y compris celles qui ne s'aligneront pas sur le génome et celles de mauvaise qualité.

Le premier fichier au format « csfasta » est un fichier fasta, dont les données de qualité sont ordonnées non pas par ligation (par cycle) mais en fonction de la séquence de code couleur. Comme décrit précédemment (voir partie 1.4.1.1), les séquences au format couleur commencent par une base correspondant à la dernière base de l'adaptateur, suivie d'une suite de chiffres entre 0 et 3 (0,1,2,3 correspondant à une des 4 couleurs des dibases). Un « . » est utilisé pour une position en code couleur dont on ne dispose d'aucune donnée (choix du fluorochrome impossible, pas de signal, ...).

Le second fichier au format « _QV.qual » comprend des scores de qualité, un pour chaque dibase lue. Ceux-ci sont calculés en utilisant le calcul de la valeur de qualité du Sanger ([Ewing & Green, 1998], [Ewing *et al.*, 1998]), le « phred », tels que :

$$QV = -10xlog10(p)$$

où QV est la valeur de qualité et p la probabilité prédite qu'une couleur détectée soit incorrecte.

Ces valeurs de qualité (QV), assignées pour chaque base, sont comprises entre 0 et 40. Les positions non renseignées des séquences (notées « . ») auront pour score -1. Ces scores sont pris en compte lors de l'alignement. Ils permettent de caractériser la qualité des séquences d'ADN et peuvent être utilisés pour comparer les efficacités des différentes technologies de séquençage.

Dans le cas du mode de séquençage paired-end, chaque fragment (notés F3 et F5) est séquencé indépendamment. On obtient ainsi un fichier csfasta et un fichier qual pour chacun des fragments, soit 4 fichiers en tout. Enfin, si l'on utilise des codes-barres, ils sont également séquencés (en début de run) afin d'attribuer chaque bille à un échantillon, et donnent eux aussi une paire de fichiers csfasta/_QV.qual.

Au cours du séquençage, le SOLiD™génère un rapport de qualité sur chaque cycle de ligation, consultable à l'aide du logiciel SETS (Figure 5.11). Ce rapport comprend plusieurs parties ou analyses, telles que : (1) la saturation du signal fluorescent pour chaque fluorochrome pour chaque quadrant de la lame (Figure 5.11 A), (2) le « satay plot » pour chaque ligation représentant le déséquilibre des couleurs et la présence de billes polyclonales (Figure 5.11 C), (3) l'auto-corrélation entre les différents échantillons. Il est également important de noter que

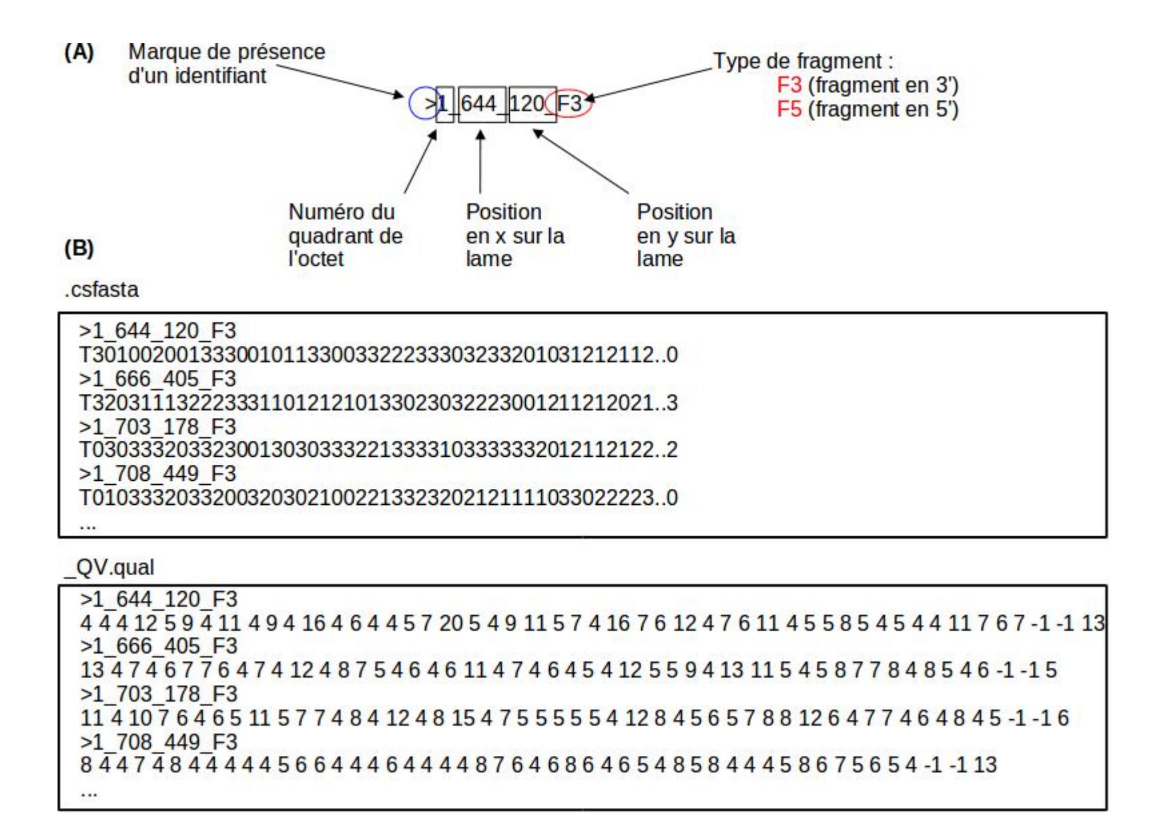


FIGURE 5.10 – Les formats fichiers bruts standards du SOLiD™ avec en (A) la notation de l'identifiant des billes et en (B) quelques lignes seulement d'un fichier « .csfasta » et « _QV.qual »

la qualité de la préparation des billes (faible proportion de billes polyclonales, absence d'un adaptateur) peut être estimée avant de passer au séquençage de tout l'échantillon. Pour cela, un run test appelé WFA pour WorkFlow Analysis est réalisé sur une faible proportion de billes.

En plus du rapport de SETS, il existe d'autres outils prenant en entrée les fichiers de séquences initiales ou bien encore les *reads* alignés, tels que par exemple le logiciel FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). Celui-ci est simple d'utilisation, rapide, et permet d'obtenir graphiquement des statistiques sur la qualité du séquençage ou des *reads* alignés (Figure 5.11).

5.3.2 Formats standards et outils de manipulation de données

En fonction du type de données générées et du niveau d'analyse (primaire, secondaire ou tertiaire), il existe des formats de fichiers de données particuliers (Tableau 5.2). Il est à noter que ces fichiers peuvent parfois avoir une taille très importante (plusieurs dizaines de Gb).

De nombreux logiciels d'alignement acceptent uniquement le format « fastq » en entrée, d'où la nécessité de créer des outils de conversion de format. Afin d'y parvenir, ou bien encore d'extraire des données ou de fournir des statistiques sur les éléments contenus, divers outils de manipulation de données ont été créés : BEDTools ([Quinlan & Hall, 2010], SAMtools ([Li *et al.*, 2009]), BamTools ([Barnett *et al.*, 2011]), picard, GATK ([McKenna *et al.*, 2010]).

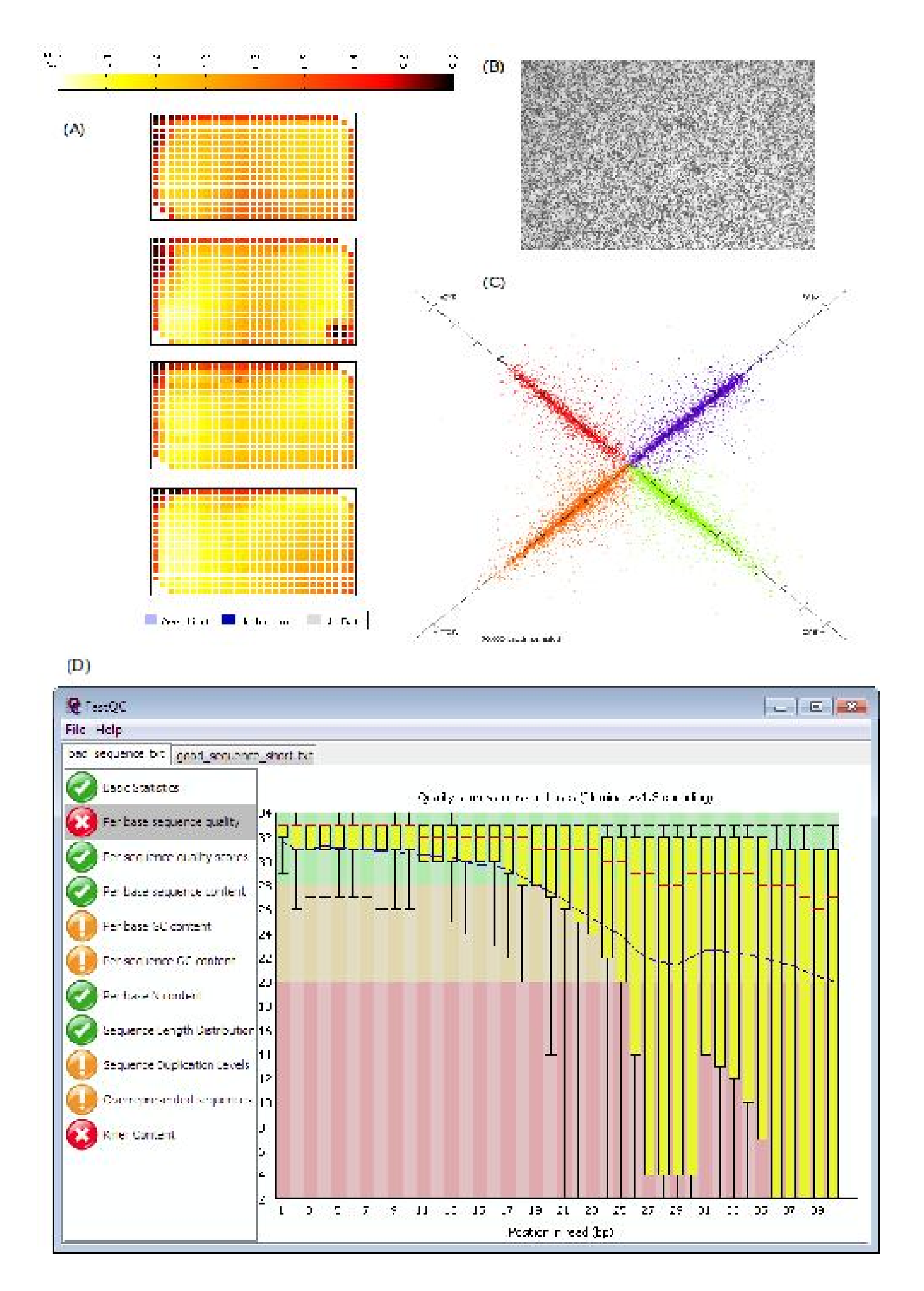


FIGURE 5.11 – Visualisation de la qualité des *reads* à l'aide des logiciels SETS ou FastQC : (A) la saturation du signal, (B) la densité de billes sur la lame avant le début du séquençage, (C) le satay plot fait sur une seule ligation, dans lequel chaque point correspond à une bille. Les points représentant des billes monoclonales sont situées sur un des 4 axes de mesure de l'intensité des 4 fluorochromes, (D) visualisation de l'interface graphique du logiciel FASTQC, permettant différentes analyses dont l'étude de la distribution des scores de qualité pour chacun des 50 nucléotides séquencés

Analyse	Format	Description
Primaire	csfasta	Format de fichier propriétaire de Life Technologies
	_QV.qual	Format de fichier propriétaire de Life Technologies
	fastq	Format principal de séquence provenant de techniques HTS
	sff	Format de fichier propriétaire de Roche
Secondaire	ma	Format de fichier propriétaire de Life Technologies
	bam	Format de fichier binaire d'alignement, converti à partir du fichier .ma
	sam	Fichier obtenu par décompression d'un fichier bam
Tertiaire	gff	Fichier d'annotation
	wig	Fichier de couverture
	bed	Fichier d'annotation
	txt	Fichier contenant des résultats ou statistiques

Table 5.2 – Les principaux formats de données du séquençage à très haut débit

Ils fonctionnent en ligne de commande, ce qui permet leur inclusion dans divers programmes et pipelines d'analyse.

Les formats de fichiers obtenus par l'analyse tertiaire dépendent de l'outil utilisé et du type d'analyse. Il n'existe pas de format standard repris par tous les logiciels ; généralement, ils utilisent en sortie des fichiers texte tabulés ou des fichiers bed pour permettre l'inclusion des résultats dans un genome browser. Il est à noter toutefois l'utilisation d'autres types de format tels que le VCF (*Variant Call Format*) pour les variants génomiques (SNP, insertions et délétions). Ce format, mis au point par le projet 1000 Génomes, est notamment utilisé par SAMtools et GATK.

5.3.3 Alignement sur le génome de référence

Le ChIP-seq est une technique de re-séquençage. Il est donc indispensable de disposer d'une référence (génome, transcrits ...) à laquelle les fragments séquencés seront comparés par alignement. Cette technique se limite donc à l'étude des organismes dont le génome a déjà été séquencé. L'alignement des fragments sur la référence, généralement un génome, est réalisé par complémentarité de bases. La version du génome utilisée est très importante pour la suite de l'analyse, et doit être choisie en fonction des besoins (annotations disponibles, qualité de l'assemblage, outils développés). En effet, entre les 2 dernières versions du génome humain, hg18 (mars 2006) et hg19 (février 2009), les principales différences sont la taille et les annotations. De plus, beaucoup d'outils même récents utilisent encore la version hg18

du génome car celle-ci dispose de plus d'annotations, que la version hg19, néanmoins plus complète. Des outils de conversion de position entre les 2 versions de génomes ont été créés tels que : liftover de ucsc ; ceux-ci permettent de convertir des annotations pour les *genome browsers* d'une version à une autre comme les positions de SNP, de motifs de fixation de facteurs de transcription (Oregano, TFBSconserved).

Cependant, ces versions de génome ne sont pas adaptées à l'étude des cellules cancéreuses, qui sont caractérisées par des amplifications et des délétions génomiques. Il est important de prendre en compte cette information pour le reste de l'étude. Il faut également tenir compte de la présence de séquences répétées (détectées grâce au logiciel repeatmasker). En effet, un fragment qui s'aligne dans une zone répétée ne peut pas être considéré comme spécifique de cette zone et ne doit donc pas être pris en compte. Au niveau de l'alignement, ces zones répétées se caractérisent par un empilement excessif de fragments sur une courte région (quelques dizaines de paires de bases). Ces empilements sont fréquemment observés dans les télomères et les centromères des chromosomes, régions riches en répétitions. L'utilisation du mode de séquençage paired-end, qui permet de s'affranchir en partie de ce genre de problème, est pour cette raison de plus en plus courante.

A l'heure actuelle, il est courant qu'une expérience de HTS génère une centaine de millions de *reads*. L'alignement d'une telle quantité de courtes séquences sur une référence ne peut se faire à l'aide des outils d'alignement classiques, tels que BLAST ou FASTA ([Altschul *et al.*, 1990]). En effet, ces derniers ont été conçus principalement pour retrouver une séquence (query) dans une référence (subject) et peuvent pour cela être paramétrés très finement afin de prendre en compte différents cas de figure (mismatches, gaps) en fonction du contexte de la recherche. Par contre, cette complexité s'accompagne d'une relative lenteur, ce qui les rend inadaptés au problème posé par le HTS. Pour cette raison, divers outils, basés sur de récents algorithmes de recherche, ont été développés afin de permettre l'alignement sur le génome de référence d'un grand nombre de courts fragments (50nt pour le SOLiD™), dans un temps raisonnable. Un logiciel, appelé mapread (Corona Lite), a été développé dans ce but par Life Technologies. Il a l'avantage de prendre en entrée les 2 fichiers issus de l'analyse primaire, à savoir le fichier csfasta (code couleur) et le fichier de qualité, ce que peu de logiciels sont encore capables de faire. D'autres logiciels pour l'alignement des séquences courtes existent comme : BOWTIE ([Langmead *et al.*, 2009]), BWA ([Li & Durbin, 2009]), BFAST ([Homer *et al.*, 2009]), ELAND, SHRIMP2 ([Rumble *et al.*, 2009], [David *et al.*, 2011]), SOAP ([Li & Homer, 2010]).

La qualité de séquençage est telle que l'on considère correctement alignés des *reads* possédant jusqu'à 2 mismatches avec la référence. Ces outils prennent également en compte la valeur de qualité de séquençage de chaque base ou dibases contenue dans les fichiers _QV.qual ou fastq, permettant d'aligner chaque base sur la référence en mettant un poids différent aux bases en fonction de leur qualité. Des jeux de séquences tests (ou benchmark) ont été créés pour permettre la comparaison de ces logiciels ([Holtgrewe *et al.*, 2011]).

Les données d'alignement sont la plupart du temps produites au format bam, devenu le format quasi-standard pour ce type de donnée. Ce format, outre le fait qu'il soit compressé et

permette de gagner de l'espace de stockage, est également indexé. L'accès aux données alignées est ainsi extrêmement rapide, ce qui permet de les parcourir de façon relativement confortable malgré leur taille (parfois plusieurs dizaines de Go). Une fois les fragments alignés, ceux-ci sont sélectionnés sur la base de leur qualité. Il peuvent être ainsi directement visualisés à l'aide d'un navigateur de génome (*genome browser*), tel que UCSC genome browser, *Integrated Genome Browser* (IGB) ou *Integrative Genomics Viewer* (IGV).

5.3.4 Recherche de pics

Le but d'une expérience de ChIP-seq est d'identifier les régions enrichies en fragments par rapport au bruit de fond et/ou à l'input. Celles-ci représentent les sites de liaison d'un facteur de transcription ou des sites étendus pour les marques d'histone.

Il convient de souligner que le ChIP est un enrichissement et non une stratégie de purification. Il faut également garder à l'esprit que certaines régions peuvent apparaître enrichies alors qu'elles n'interagissent pas avec la protéine d'intérêt, par exemple les régions du génome prédisposées à la fragmentation. Cela est probablement influencé par des facteurs tels que les éléments répétitifs et le niveau d'ouverture de la chromatine. De plus, le génome de référence utilisé et la variation du nombre de copies de certaines régions chromosomiques, appelées amplicons dans les cancers, conduisent également vers la génération de faux positifs. Il faut donc bien prendre en compte la distribution théorique des *reads* dans ces régions, notamment le déséquilibre de brin, pour filtrer les artefacts (Figure 5.12).

Avant de lancer un programme de recherche de pics, on filtre généralement les données pour supprimer les artefacts tels que les amplifications PCR. Ce filtre est maintenant inclus dans la plupart des algorithmes.

De très intéressantes revues ont été réalisées par Pepke, Wilbanks et leurs collaborateurs sur la comparaison des algorithmes de recherche de pics pour les facteurs de transcription ([Pepke *et al.*, 2009], [Wilbanks & Facciotti, 2010]). Ces algorithmes ne prennent pas les mêmes paramètres en entrée et ne donnent pas forcément des résultats similaires en terme de taille des sites de fixation et de quantité (Figure 5.13). Le choix de l'algorithme à utiliser est dicté par le type de facteur de transcription étudié. Ainsi pour un facteur de transcription site spécifique, on préfèrera un logiciel donnant des pics étroits. En revanche, pour des facteurs tels que Cbp (Creb-Binding Protein), ou son homologue p300, qui est un co-activateur pour un grand nombre de facteurs de transcription, notamment Creb, E2F, Jun, Fos, on cherchera des régions de taille plus étendue comme pour les marques d'histones éventuellement. Un alignement test a également été créé pour évaluer les algorithmes. Celui-ci correspond à un alignement disposant de pics et également d'artefacts tels que des empilements (Figure 5.12).

Une fois les pics identifiés, ils peuvent être inclus et visualisés, en plus des alignements, à l'aide d'un genome browser, ce qui permet de juger de leur qualité.

Enfin, ces informations sont maintenant collectées dans des outils ou bases de données

dédiés tels que Epigraph (http://epigraph.mpi-inf.mpg.de/WebGRAPH/).

5.3.5 Découverte et recherche de motifs

Une fois localisés les pics potentiels de fixation du facteur de transcription, on peut rechercher la présence de motifs de fixation de facteurs de transcription dans les séquences présentes sous les pics. On peut ainsi conforter les résultats obtenus pour les facteurs de transcription interagissant directement avec l'ADN, et mettre en exergue la présence de co-facteurs dans le cas de p300 par exemple.

De plus, à partir d'une liste de pics de bonne qualité, il est également possible d'améliorer les motifs par la constitution d'une collection de séquence (type de représentation des motifs) : par exemple, celles contenues dans les bases de données de JASPAR ([Wasserman & Sandelin, 2004]) et TRANSFAC ([Matys *et al.*, 2003]), ou bien encore UniPROBE pour la souris. Elles collectent des informations sur les sites potentiels de liaison dans une matrice poids-positions ou logo.

Divers outils de recherche de motifs ont ainsi été adaptés aux données issues du ChIP-seq, tels que RSATools avec peakmotifs, MEME avec MEME-ChIP ([Machanick & Bailey, 2011], DREME ([Bailey, 2011]). Certains pipelines d'analyse comme la librairie rGADEM permettent, à l'aide de plusieurs librairies R, de rechercher les motifs à partir des pics déterminés par PICS (Figure 5.13). Ces outils nous renseignent sur la probabilité d'obtenir une liste enrichie en sites potentiels de fixation d'un facteur, en fonction de son occurrence sur le génome. Le problème est la détermination du modèle de fond et des séquences finalement utilisées pour la recherche de motifs. MEME-ChIP n'utilise que les meilleures séquences pour construire le motif.

Ces outils proposent généralement l'utilisation de deux jeux de séquences : les séquences sous les pics provenant du ChIP et celles provenant de la condition contrôle (input).

Les pics, mais aussi les *reads*, peuvent être visualisés sur le génome à l'aide de genome browser qui permet l'ajout d'annotations (track) telles que : les séquences répétées obtenues par repeatmasker, des données de transcriptome, de CGH, afin de pouvoir mieux interpréter les données.

Un autre critère important pour la recherche de motifs, est leur conservation au cours de l'évolution ([Cai *et al.*, 2010]). De nombreux outils de recherche de motifs utilisent également ces données de conservation pour affiner leur analyse (comme ECRbase [Loots & Ovcharenko, 2007] par exemple).

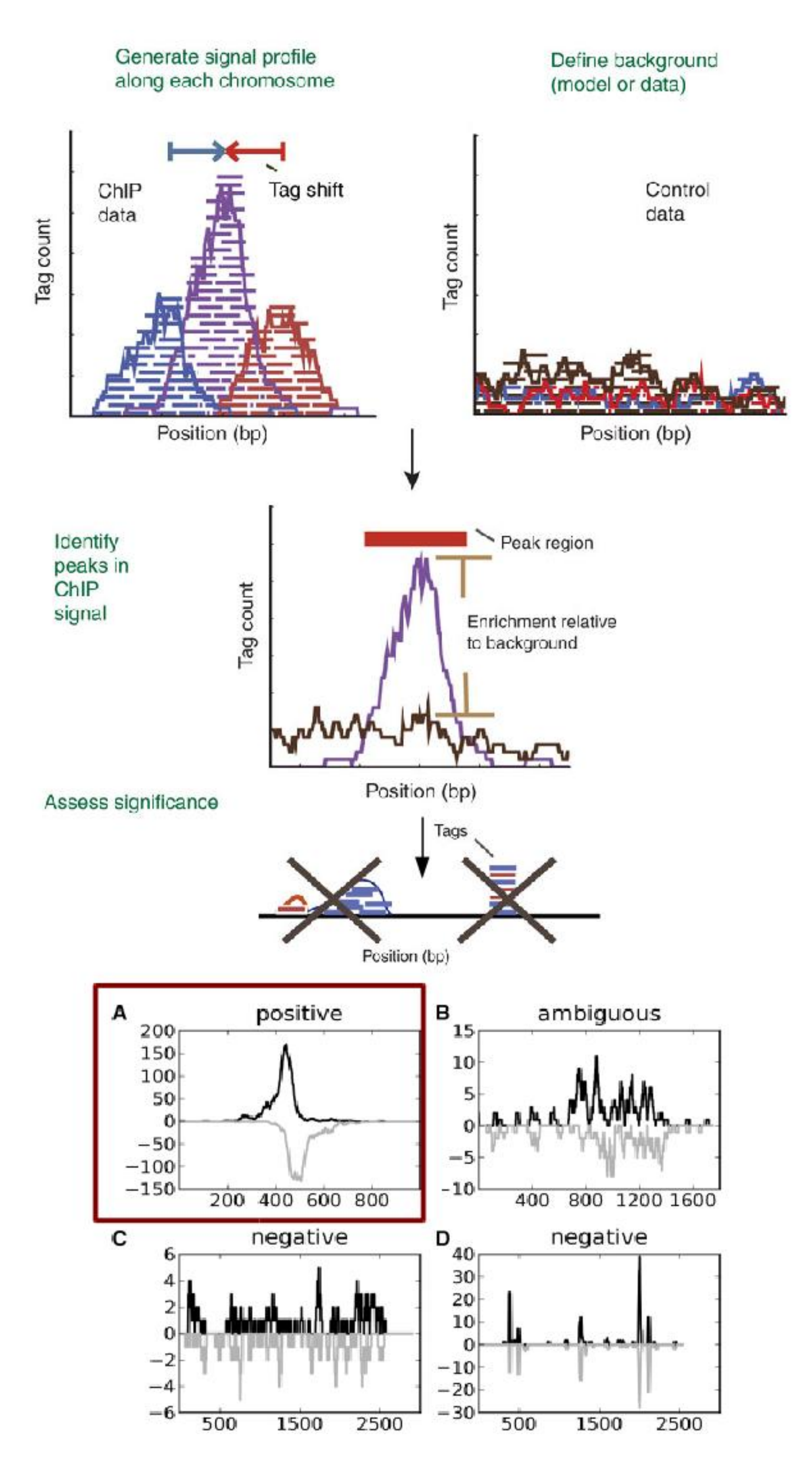
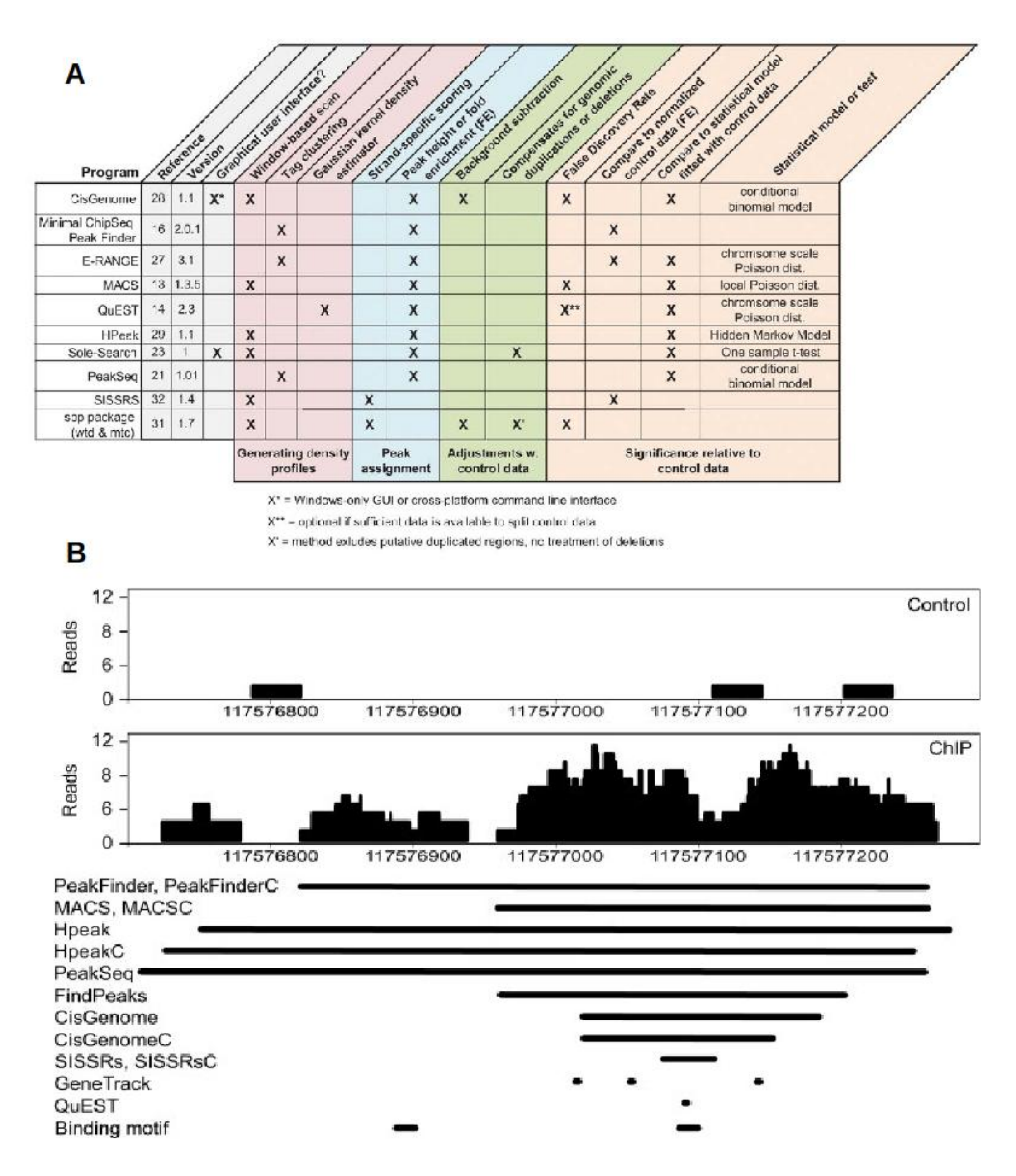


FIGURE 5.12 – Choix de la méthode de détection de pics et représentation des artefacts (Adapté de [Pepke *et al.*, 2009] et [Rye *et al.*, 2011])


A

Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromsome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromsome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
sop package (wtd & mtc)	31	1.7		X			X		X	X'	X			
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			

X* = Windows-only GUI or cross-platform command line interface
X** – optional if sufficient data is available to split control data
X' = method exludes putative duplicated regions, no treatment of deletions

Figure 5.13 – Les différents logiciels de recherche de pics (A) Tableau récapitulatif des principales méthodes (B) Représentation des pics obtenus par ces différentes méthodes à une position donnée du génome (Adapté de [Wilbanks & Facciotti, 2010])

5.3.6 Annotation et visualisation des résultats

A partir d'une liste de pics, plus ou moins fiables en fonction du score de qualité qui leur est associé, on cherche à savoir quels sont les gènes potentiellement régulés par le facteur étudié, ceci dans le but de créer des réseaux de gènes en évaluant, comme pour les puces à ADN, l'enrichissement fonctionnel de cette liste de gènes cibles en un processus, une voie de signalisation.

La pratique standard pour associer les pics à un gène est d'utiliser certains critères comme la distance au site d'initiation de la transcription (ou TSS) ou à un élément structural du gène (intron, exon). Par exemple, Johnson et collègues (2007) ont cartographié les pics à ± 20 kb du TSS d'un gène, alors que Wederell et collègues (2008) ont utilisé une distance comprise entre -10 kb du TSS et +1 kb du site de terminaison de la transcription. Chen et collègues (2008) ont utilisé une méthode plus sophistiquée, et déterminé la répartition des distances des pics aux TSS des gènes pour chaque facteur évalué. Les pics sont ensuite associés au gène le plus proche, à l'intérieur de cette répartition. Cependant, il est à noter que dans le cas des enhancers (1) le gène le plus proche n'est pas forcément celui qui sera régulé par le facteur de transcription étudié, et (2) un facteur de transcription peut réguler plusieurs des gènes qui l'entourent. Ainsi le 3C-seq, le ChIA-PET et d'autres extensions de ces approches permettent de détecter les interactions entre les locus génomiques et d'identifier les sites de régulation associés au TSS lors de l'initiation de la transcription pour aider à résoudre ces problèmes ([Dekker *et al.*, 2002], Dostie et al. 2006, Simonis et al. 2006, Zhao et al. 2006). Divers outils d'annotations ont été implémentés : Savant ([Fiume *et al.*, 2010]), ChIPpeakAnno ([Zhu *et al.*, 2010]), CEAS ([Shin *et al.*, 2009]), cisgenome ([Barozzi *et al.*, 2011]) et GREATER ([McLean *et al.*, 2010]).

Chez les mammifères, près de la moitié des sites de liaison identifiés sont associés à des gènes inactifs ([Hatzis *et al.*, 2008]). En effet, de nombreux facteurs de transcription sont soit des cofacteurs soit des répresseurs. Il est donc important de savoir si un gène est bien fonctionnellement lié au facteur de transcription auquel il est associé. Pour fournir un argument prouvant que les sites sont fonctionnels, plusieurs méthodes peuvent être utilisées comme : (1) étudier l'expression différentielle des gènes avec et sans site de fixation associé ([Johnson *et al.*, 2007], [Chen *et al.*, 2008]), (2) évaluer l'expression de gènes cibles, dans les cellules dans lesquelles l'expression du facteur d'intérêt a été réduite ou supprimée, et (3) rechercher la concordance des sites de liaison et des modifications des histones qui délimitent les régions enhancer et promoteur, comme H3K4me1 et H3K4me3 ([Barski *et al.*, 2007]).

5.3.7 Bases de données dédiées aux données HTS

Le NCBI et l'EBI ont également mis en place des bases de données pour accéder aux données de séquençage à très haut débit : Sequence Read Archive (SRA) et European Nucleotide Archive (ENA), respectivement. Mais avec l'explosion de ces techniques, le nombre d'expériences croît de manière exponentielle, ce qui nécessite l'augmentation des capacités de stockage ; là où quelques dizaines de Mb suffisaient pour les puces à ADN, il faut maintenant plusieurs centaines de Gb de stockage pour une expérience de HTS. En effet, ces bases de données permettent (1) le stockage des fichiers bruts (csfasta/qual pour SOLiD ou fastq)

qui représentent plusieurs Gb, (2) des alignements (fichiers bam) et (3) de données résultant de pipeline d'analyses tertiaires parfois (format bed, pour la localisation des pics provenant d'expérience de ChIP-seq).

De la même manière que MIAME pour les données de puces à ADN, le FGED a mis au point le Minimum Information about a highthroughput Nucleotide SeQuencing Experiment (MINSEQ) pour les données de séquençage à très haut débit.

5.4 Elaboration d'outils et de méthodes d'analyse pour les données de ChIP-seq

Afin de pouvoir analyser les données de ChIP-seq issues du séquenceur à très haut débit SOLiD™présent sur la plateforme TGML, un pipeline de traitement des données de ChIP-seq a été développé. Il effectue les analyses secondaires et tertiaires des données de ChIP-seq réalisées en mode de séquençage fragment, mais également paired-end, l'analyse primaire étant réalisée sur le cluster du séquenceur. Ce développement récent, n'a donné lieu à aucune publication. Cependant il est utilisé en routine sur la plateforme TGML et utilisé dans le cadre de collaborations. Je l'ai également utilisé pour analyser des données de ChIP-seq, expérience que j'ai moi-même réalisé. Ces résultats ne seront toutefois pas montré dans ce manuscrit.

Ce pipeline étant principalement destiné à l'analyse de données provenant de la plateforme, il a été construit sur la base des outils proposés par Applied Biosystems (Bioscope, Corona lite). Toujours en développement, afin de s'adapter aux besoins de la plateforme, il est écrit en langage bash, particulièrement approprié pour l'intégration de logiciels d'origine diverse. Ainsi, il intègre différents outils publics, ainsi que des scripts et programmes développés au laboratoire. En effet, l'évolution technique des séquenceurs à très haut débit étant très rapide (voir Chapitre 1 1.1). Ainsi depuis 2009 nous avons changé plusieurs fois de version du SOLiD™passant d'une version v3.0 à la version v3.5 en 2010 et enfin à la v4 en 2011. Ceci a entraîné de nombreux changement tant au niveau expériemental (longueur des *reads*, taille de sonication) qu'au niveau informatique (version de bioscope, modification des programmes et pipelines).

Actuellement, pour un échantillon, on dispose en général d'environ 35-40 millions de billes et donc autant de *reads*. Le pipeline met environ 6 heures par échantillon pour les analyses secondaires et tertiaires. Un fichier de log permet de suivre la progression des différentes étapes et le comportement des outils dont les sorties sont redirigées dans ce fichier.

5.4.1 Choix des logiciels et stratégies

Les données de séquençage du SOLiD™provenant de l'analyse primaire étant dans les formats « csfasta » et « _QV.qual », le choix d'utiliser le logiciel d'alignement fourni par la suite bioscope appelé mapread s'est naturellement imposé. Toutefois, les logiciels BOWTIE et BWA ont été testés, avec conversions pour la prise en charge de ces formats, sans montrer

de différences de qualité avec le logiciel mapread de bioscope. De plus, par défaut ces autres outils ne parallélisent par leurs tâches d’alignement comme le fait la suite bioscope et donc ils se révèlent nettement plus lents, sans apporter de réel avantage.
Le fichier de résultats de l’alignement étant au format compressé « bam », sa conversion à l’aide de la suite d’outils samtools permet l’obtention d’un fichier plat au format « sam » plus adapté aux traitements. Ce dernier est ensuite utilisé au niveau de l’analyse tertiaire développée en fonction des particularités de l’analyse de données de ChIP-seq présentées dans la partie précédente de ce chapitre.

L’étape suivante consiste à réaliser la détection de pics. Quelques outils ont été testés avec différents paramètres : MACS, Hpeaks, MICSA. Le choix s’est porté finalement sur l’intégration de MACS, car celui-ci donne des pics plus étroits que Hpeaks. Cependant, il génère beaucoup d’artefacts (empilement de *reads* anormalement pris en compte) dus à une conception d’un modèle de pics assez difficile. C’est la raison pour laquelle, en parallèle à MACS, un autre outil de détection de pics développé au laboratoire et nommé Picor a été intégré au pipeline (voir ci-après).

Les pics obtenus, sont ensuite filtrés pour ne conserver que ceux qui ne chevauchent pas une région répétée. Il en résulte un fichier de résultat au format « bed » servant à (1) l’analyse fonctionnelle par divers scripts implémentés sur la plateforme, (2) la recherche et/ou découverte de motifs, après récupération des séquences fasta sous les pics, à l’aide de peakmotifs de la suite logicielle RSATools installée sur un serveur de notre laboratoire et (3) la visualisation des pics avec un navigateur de génome comme IGV ou UCSC en parallèle des alignements au format « bam ».

On peut ainsi à partir de la localisation des pics déterminer quels sont les gènes cibles potentiellement régulés par le facteur de transcription étudié et analyser l’enrichissement fonctionnel de ceux-ci à l’aide d’outils comme GREATER ou DAVID knowledgebase, ceci afin de construire le réseau de gènes régulés par le facteur de transcription. La distribution de la localisation des pics est également étudiée sur le génome et au niveau de la structure des gènes.

Enfin un contrôle qualité des données avant et après alignement a été ajouté au pipeline à l’aide du logiciel FastQC, celui-ci venant en complément du rapport de séquençage fourni par SETS.

5.4.2 Picor : un nouvel outil pour la recherche de pics

Face au problème de détection de pics (voir partie 5.3.4) et vus les résultats donnés par certains algorithmes sur nos données, des chercheurs du laboratoire ont conçu un nouvel algorithme pour la détection de sites de fixation de facteurs de transcription à partir de données de ChIP-seq, nommé Picor (Figures 5.14 et 5.15). Cet algorithme non publié à ce jour a été intégré au pipeline d’analyse des données en parallèle d’un autre outil, MACS. J’ai testé et intégré ce nouvel outil dont je n’ai pas pris part à la conception, à mon pipeline.

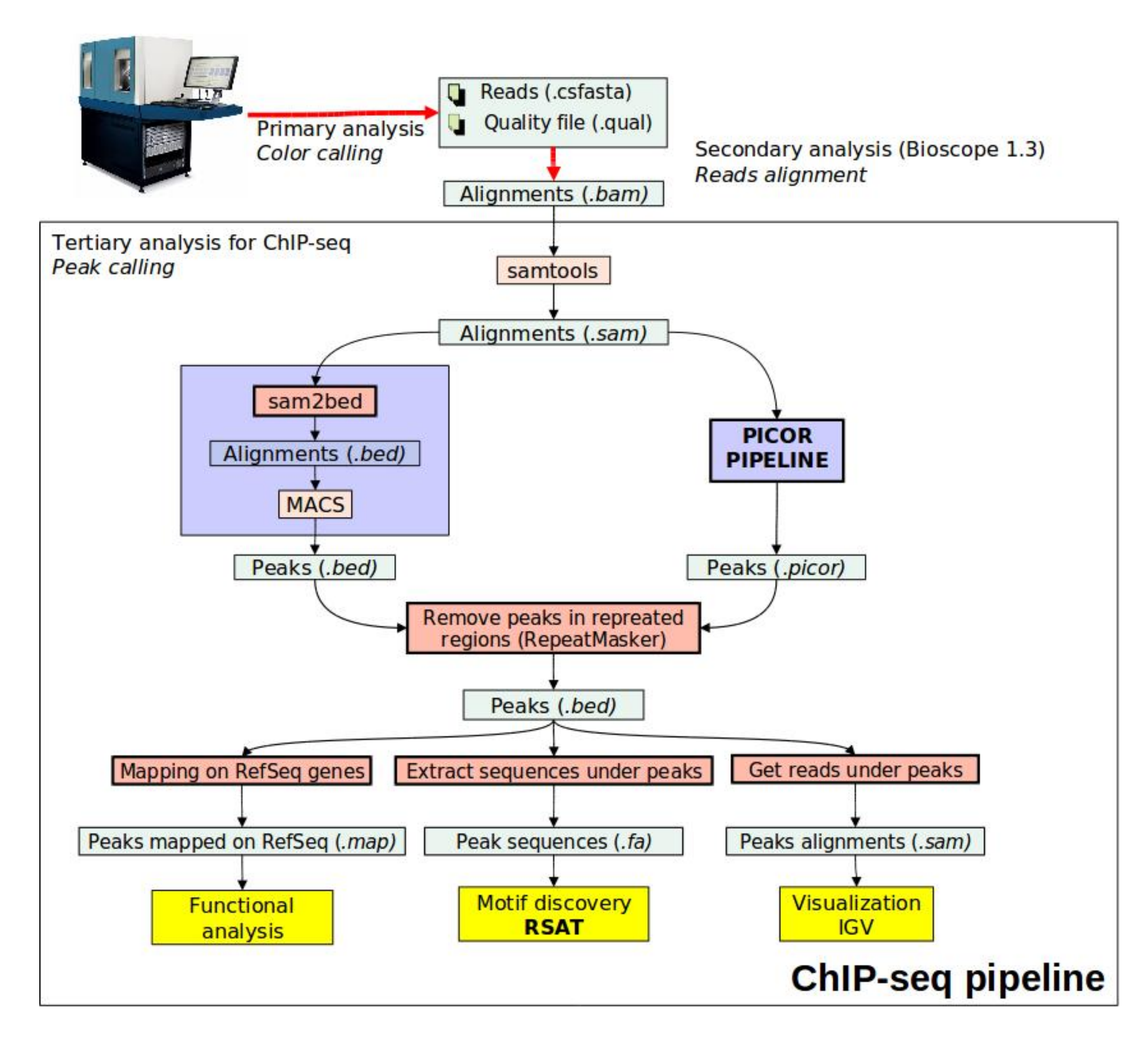


Figure 5.14 – Principales étapes du pipeline d'analyse des données de Chromatine ImmunoPrécipitation (ChIP-seq) sur la plateforme TGML

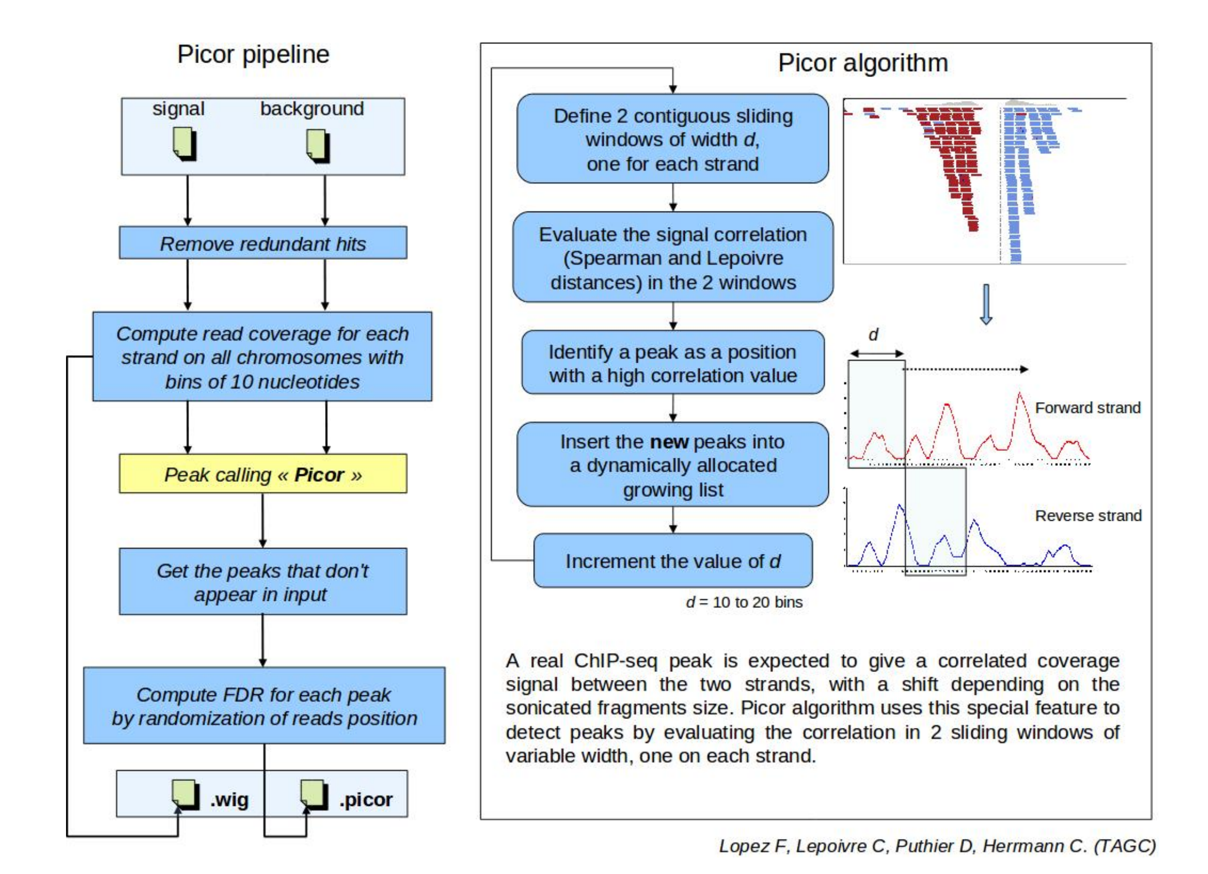


FIGURE 5.15 – Principe de l'algorithme et pipeline d'analyse du programme de détection de pics développé au laboratoire TAGC

Cet outil prend en entrée deux fichiers d'alignement correspondant au ChIP-seq et à l'input, et dont les *reads* redondants ont été supprimés car ils sont généralement dus à des amplifications PCR ou génomiques. Il génère ensuite un fichier wig par chromosome contenant le nombre de *reads* tous les 10 nucléotides, sur chaque brin séparément. Ce sont ces fichiers wig que Picor analyse pour trouver les pics.

Le principe de l'algorithme de Picor part du postulat que pour un facteur de transcription séquence spécifique, on doit observer au niveau des données un déséquilibre de liaison. Si on étudie la couverture des deux brins du génome à l'aide de fenêtres glissantes de taille variable mais distantes d'une distance d correspondant à la moitié de la taille de sonication, on doit donc obtenir un pic sur chaque brin tout d'abord sur le brin + puis sur le brin - (Figure 5.15).

Il en résulte en sortie un fichier de type bed donnant la localisation des pics dépassant un certain FDR calculé pour chaque pic et comportant en plus de la localisation à la base près des pics, la taille de la fenêtre et la distance de corrélation de Spearman.

5.5 Analyse de données en collaborations

Le pipeline est utilisé en routine sur la plateforme TGML pour l'analyse des données du séquenceur SOLiD^{TM}. J'ai ainsi réalisé l'analyse des données d'expériences de ChIP-seq entièrement réalisées au niveau de la plateforme TGML, mais provenant de projets collaboratifs avec d'autres laboratoires de recherche :

(1) la collaboration avec les Docteurs Max Chaffanet et Christophe Ginestier de l'équipe du Dr Daniel BirnBaum du Centre de Recherche en Cancérologie de Marseille (CRCM) porte sur la définition des cibles du facteur de transcription ZNF703 dans le cancer du sein ;

(2) la collaboration avec Nathalie Sakakini, étudiante en seconde année de thèse au TAGC sous la co-direction des Docteurs Jean Imbert et Thierry Virolle de l'unité Inserm U898 « stem cells, development and cancer » de Nice porte sur l'étude de la fixation des facteurs de transcription EGR1 et B-Caténine dans deux lignées de Glioblastome ;

(3) la collaboration avec le Dr Salvatore Spicuglia, faisant alors partie de l'équipe du Dr Pierre Ferrier au Centre d'Immunologie de Marseille-Luminy (CIML) portant sur l'étude du facteur de transcription TLX3 lors du développement des lymphocytes T chez la souris ;

(4) la collaboration avec le Dr Saadi Khochbin, directeur de l'équipe « épigénétique et signalisation cellulaire » de l'Institut Albert Bonniot de Grenoble, portant sur l'analyse de la localisation du variant d'histone tH2B lors de la spermatogénèse chez la souris.

Dans ces collaborations, mon travail a consisté en l'alignement des données brutes de séquençage, en la vérification de la qualité du séquençage, à la détection de pics et pour la plupart des cas à l'analyse de ceux-ci (statistique de distribution, annotation, recherche de motifs). Ces collaborations n'ont pas donné lieu pour le moment à des publications.

Enfin, ce pipeline a été en partie utilisé pour générer des données à partir d'une nouvelle approche, appélée Mnase-Cap développée en collaboration avec le Dr Salvatore Spicuglia. Cette approche est une technique combinant celle du Mnase-seq et de la capture sur lame (ou en solution sur bille magnétique) de région génomique ciblée. Ceci permet d'augmenter la couverture des régions ciblées et de mieux étudier le positionnement des nucléosomes au niveau des régions régulatrices de gènes d'intérêt.

Une publication de cette nouvelle approche du positionnement des nucléosomes est en cours de préparation.

5.6 Discussion et perspectives

Beaucoup de questions restent en suspens

Le séquençage à très haut débit n'est pas encore une technologie ne soulevant aucun problème technique. Il reste en effet pas mal de questions laissées en suspend et qui font l'objet de conférences internationnales, comme les puces à ADN à une autre époque.

En effet, la communauté scientifique n'a pas encore fixée de protocole spécifique du ChIP-seq concernant l'utilisation du mode fragment par rapport au paired-end, la longueur des reads à utiliser pour le ChIP-seq, la taille optimale des sonications, les meilleures méthodes pour la détection de pics, ou bien encore l'impact des amplifications génomiques sur les données de ChIP-seq.

Amplification dans les cancers

Les analyses que j'ai pu réaliser sur des échantillons de ChIP-seq, obtenus à partir de lignées cancéreuses (cancers du sein ou glioblastomes), ont montré que le nombre de pics détecté dans les régions amplifiés est supérieur aux autres régions du génome.

En effet,l'amplification génomique peut être défini comme un processus génétique conduisant à la multiplication sélective du nombre d'exemplaires d'un gène (ou d'un groupe limité de gènes adjacents, définissant un « amplicon »), ce qui participe à l'oncogenèse dans plusieurs types tumoraux (http://www.sanger.ac.uk/genetics/CGP/Census/amplification.shtml).

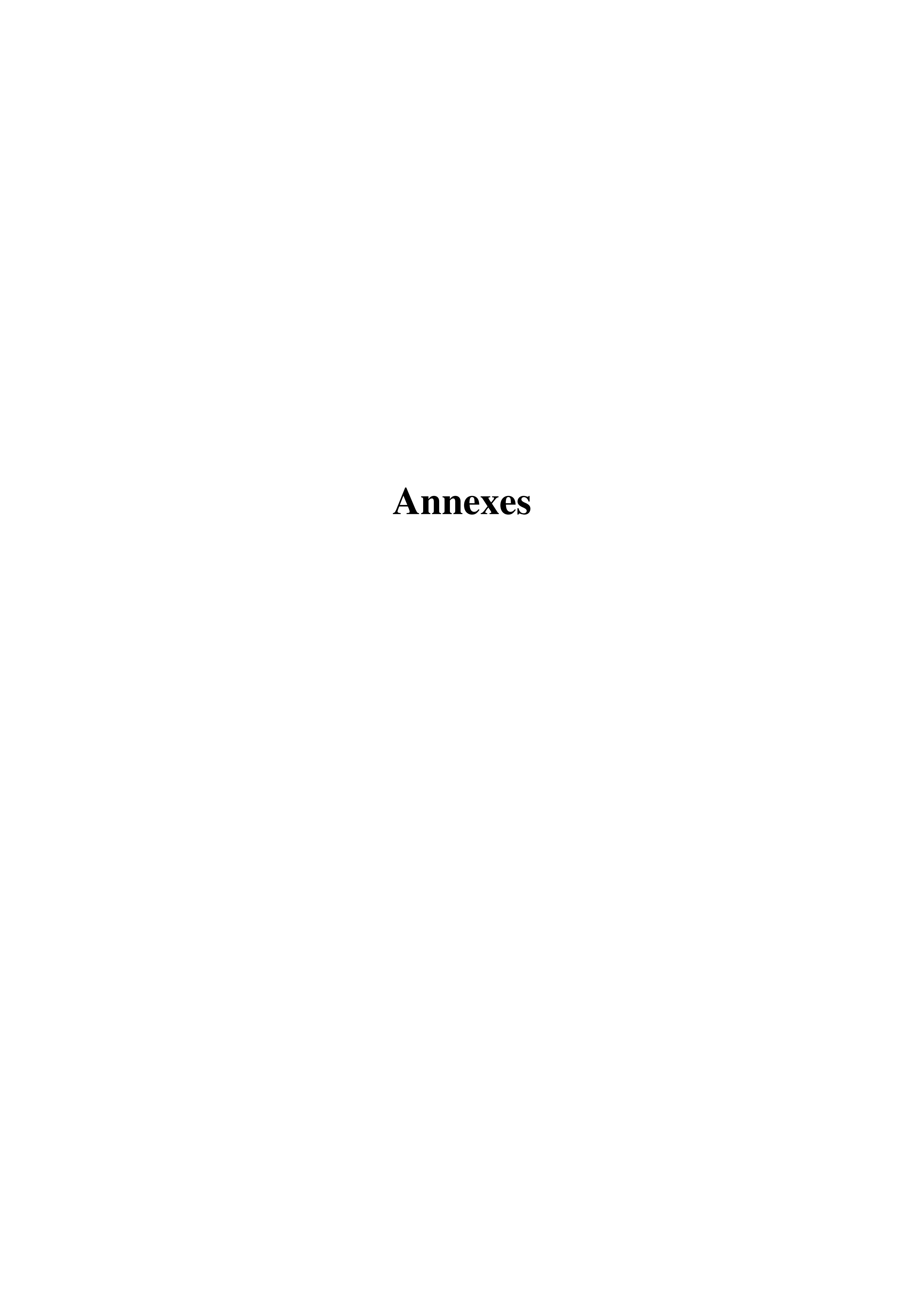
Reste à savoir si cela est dû à l'amplification génomique, qui enrichit artificiellement le nombre de fragments aléatoirement immunoprécipité et correspondant au bruit de fond, ou à des régions régulatrices importantes.

Afin de répondre à cette question, on s'oriente peu à peu vers le reséquençage du génome des tumeurs [Ross & Cronin, 2011]. En effet, chaque lignée cancéreuses possèdent sa propre amplification différente en nombre de copie et en régions génomiques d'une autre [Stephens *et al.*, 2009].

Perspectives de développement

Le développement du pipeline va être poursuivie dans le cadre d'un poste sur la plateforme. En effet, ce pipeline utilisé en interne nécessite une homogénéisation avec l'autre pipeline développé sur la plateforme et permettant la détection des SNP et *small indels*. Ces deux pipelines ont déjà en commun l'analyse secondaire, reste à les intégrer dans une interface graphique conviviale pour une utilisation externe. La laboratoire souhaite à ce sujet installé sur un serveur une version locale de Galaxy [Giardine *et al.*, 2005], avec des ressources communes comme la suite RSATools, l'outil de détection de pics PICOR, ou bien encore les différents pipelines d'analyses de la plateforme TGML.

Afin de pouvoir en partie poursuivre mes travaux de thèse, et d'en apprendre encore d'avantage sur ces mécanismes complexes mais très intéressants de la régulation de l'expression des gènes, je vais être embauchée en tant qu'ingénieur de recherche en CDD au niveau de la plateforme TGML. Mon rôle sera d'améliorer le pipeline ChIP-seq, mais surtout de mettre en place l'analyse complexe des données de RNA-seq pour laquelle rien n'est disponible au laboratoire pour le moment.



Annexes

ANNEXE A

Manuel d'utilisation de la librairie R : AgiND

The AgiND package

Aurélie Bergon and Denis Puthier[1]

July 30, 2007

1. INSERM TAGC/ERM206, Parc scientifique de Luminy case 928, MARSEILLE, FRANCE
`puthier@tagc.univ-mrs.fr`
http://tagc.univ-mrs.fr

Contents

1 Introduction

This document is intented to provide a brief overview of the **AgiND** package. The library was developped for diagnosis and normalization of one channel and two channel Agilent microarrays. Note, that the AgiND requires at least R-2.5.0 version.
Briefly, the AgiND package contains high level function for:

- diagnosis (Boxplot, color-coded images and MA plot) for both gProcessedSignal (rProcessedSignal) and gMeanSignal (rMeanSignal) intensities;
- normalization (lowess or quantiles method);
- convertion of an AgilentNorm or an AgilentNormRG object into an ExpressionSet object to ensure co-patibility with other bioconductor packages.

2 Getting started

2.1 Load the AgiND library

- After installation, load the AgiND package using the /library/Fonctions:

```
> library(AgiND)
```

- Exemple data are located in the R installation path of the AgiData package and for the need of this demonstration, user should change its working directory using the following command:

```
> library(AgiData)
> setwd(system.file(".", package = "AgiData"))
```

Note: If a MIAME and phenoData files are provided they should be located in the ExpData directory and named ”miame” and ”phenodata.txt” respectively.
Type the following command to acces help:

```
> help(AgiND)
> help.search("AgiND")
```

2.2 Note about quantification files

The library supports quantification files (with ”txt” extension) derived from the Agilent Feature Extraction software.
As user way perform both one channel (one sample) and two channel (two samples) hybridization, the AgiND package was developped to handle both approaches. For data acquisition, user shoul use the getAgilentBatch function which uses read.MIAME, read.phenoData, readAgilent and checkDim functions. In the case of two colors approach, user must set the RG argument to TRUE.

```
> args(getAgilentBatch)

function (n = NULL, RG = FALSE, path = ".", recursive = FALSE,
    flag = 2)
NULL
```

Numero	Letter	Flag name	Flag description
1	a	gIsSaturated	Feature is saturated
1	b	rIsSaturated	Feature is saturated
2	c	gIsFeatNonUnifOL	Feature is not uniform
2	d	rIsFeatNonUnifOL	Feature is not uniform
3	e	gUsPosAndSignif	Feature is not positive and significant
3	f	rIsPosAdSignif	Feature is not positive and significant
4	g	gIsFeatPopnOL	Feature is a population outlier
4	h	rIsFeatPopnOL	Feature is a population outlier
5	i	IsManualFlag	Feature is manually marked
6	j	gIsBGNonUnifOL	Background is not uniform
6	k	rIsBGNonUnifOL	Background is not uniform
7	l	gIsBGPopnOL	Background reading is population outlier
7	m	rIsBGPopnOL	Background reading is population outlier

Table 1: Table of the different flags.

2.2.1 One-channel hybridizations

As in the case of two channel hybridization the getAgilentBatch function will extract information from user-provided files and return an instence of class AgilentBatch.

```
> myob <- getAgilentBatch(1:4, path = "OneColor/", flag = 1:7)

        There are  4  files in the working directory:
         /home/aurelie/R/i686-pc-linux-gnu-library/2.5/AgiData/OneColor
        Reading all files
                 file 1 --->  US45102986_251487911262_S01_GE1-v5_95_Feb07_1_1.txt
                 file 2 --->  US45102986_251487911262_S01_GE1-v5_95_Feb07_1_2.txt
                 file 3 --->  US45102986_251487911262_S01_GE1-v5_95_Feb07_1_3.txt
                 file 4 --->  US45102986_251487911262_S01_GE1-v5_95_Feb07_1_4.txt

--> Creating an object of class AgilentBatch

                An object of class AgilentBatch (one color)
                Memory used:  16717708
                Number of samples= 4
                Number of spots=  45018
```

```
Dimensions of arrays=  532 rows x  85 columns
This object contains the following informations:
         -  gP
         -  gBGM
         -  gM
         -  fileNames
         -  PosX
         -  PosY
         -  CtrT
         -  PN
         -  GN
         -  SN
         -  Desc
         -  PhenoD
         -  Miame
         -  Flag
         -  Row
         -  Col
```

Informations about object size can be obtained using the following commands:

```
> ncol(myob)
> nrow(myob)
> dim(myob)
> length(myob)
```

2.2.2 Two-channels hybridizations

As in the case of single channel approach, quantification files should be derived from the same microarray plateform and thus contain the same number of elements.

```
> myobRG <- getAgilentBatch(1, RG = TRUE, flag = 1:7, path = "TwoColors/")
```

When RG is set to TRUE getAgilentBatch will construct an AgilentBatchRG object which differs slightly from the AgilentBatch object since it will contain rP, rM and rBGM slots.

2.3 Building phenoData and MIAME files

Although the easiest way to create phenoData and Miame information is to provide a well-formatted file in the ExpData Directory, user may use alternative solution. Indeed, user may fill the slots by invoking:

```
> dataPhenoD <- data.frame(x = 1:4, y = rep(c("Brain", "Heart"),
+     2), z = I(LETTERS[1:4]), row.names = paste("Sample", 1:4,
+     sep = "_"))
> metaData <- data.frame(labelDescription = c("Numbers", "Tissue",
+     "Condition"))
```

```
> PhenoD(myob) <- new("AnnotatedDataFrame", data = dataPhenoD,
+     varMetadata = metaData)
> PhenoD(myob)

  rowNames: Sample_1, Sample_2, Sample_3, Sample_4
  varLabels and varMetadata:
    x: Numbers
    y: Tissue
    z: Condition

> Miame(myob) <- new("MIAME", title = "There is an exemple of MIAME file",
+     name = "Experience name", lab = "INSERM TAGC/ERM206", contact = "Mr Dupond",
+     url = "http://tagc.univ-mrs.fr", abstract = "an abstract describing the experiment")
> Miame(myob)

Experiment data
  Experimenter name: Experience name
  Laboratory: INSERM TAGC/ERM206
  Contact information: Mr Dupond
  Title: There is an exemple of MIAME file
  URL: http://tagc.univ-mrs.fr
  PMIDs:

  Abstract: A 5 word abstract is available. Use 'abstract' method.
```

Futhermore,user can call the read.AnnotatedDataFrame and read.Miame functions.

```
> PhenoD(myob) <- read.AnnotatedDataFrame(filename = paste(getwd(),
+     "/phenoData.txt", sep = ""), sep = "\t", head = T, fill = NA,
+     quote = "")

> Miame(myob) <- read.MIAME(filename = paste(getwd(), "/miame",
+     sep = ""))
```

2.4 Object informations

2.4.1 Class description

Complete description of slots for classes AgilentBatch and AgilentBatchRG is provided in the Table 2.

2.4.2 Accessing slots

Different components or slots of the microarray may be accessed using the @ operator or alternatively using the slot function.
object@slot.name
slot(object,"slot.name")

Slot name	Column name of quantification file	Description
gP	gProcessedSignal	Matrix of the normalized signal obtained by the Feature Extraction software on the green chanel.
gM	gMeanSignal	Matrix of the mean signal measured in the green channel.
gBGM	gBGMeanSignal	Matrix of the mean background signal measured in the green channel.
rP	rProcessedSignal	Matrix of the normalized signal obtained by the Feature Extraction software in the red channel (for AgilentBatchRG and AgilentNormRG class object).
rM	rMeanSignal	Matrix of the mean signal measured in the red channel (for AgilentBatchRG and AgilentNormRG class object).
rBGM	rBGMeanSignal	Matrix of the mean background signal measured in the red chanel (for AgilentBatchRG and AgilentNormRG class objects).
fileNames		Vector containing names of the files used to build the AgilentBatch or AgilentBatchRG object.
PosX	Col	Vector of the column localization of the spot on array.
PosY	Row	Vector of the row localization of the spot on array.
Desc	Description	Vector containing probe annotation.
GN	GeneName	Vector containing gene names for corresponding probes.
PN	ProbeName	Vector of the probe name.
SN	SystematicName	Vector of the sustematic name of the gene corresponding to the probe.
Flag	gIsFeatureNonUnifOL	Matrix which allows to know if it's a qood quality spot (feature is not uniform).
CtrT	ControlType	Vector of the control type: -1 (negative control),0 (sample) ou 1 (positive control)
PhenoD		The phenodata.txt file is in the ExpData directory of the working directory. This slot is an phenoData class object.
Miame		The miame.txt file contain Minimum Information About Microarray Experiment and is in the ExpData directory of the working directory. This slot is an MIAME class object.
Row		Vector of the number of array row.
Col		Vector of the number of array column.

Table 2: Table of the different slot contained in an AgilentBatch or AgilentBatchRG class object.

If implemented, user may also use the corresponding method.
slot.name(object)
For a slot containing a matrix, the Following command will be valid.
object@slot.name[i,j]
slot(object,"slot.name")[i,j]
slot.names(object)[i,j]
object[i,j]

where i correspond to one or several spots
and j correspond to one or several arrays

Exemples for matrix:

```
> myob@gP[1:20, 1:2]
> slot(myob, "gP")[1:20, 1:2]
> gP(myob)[1:20, 1:2]
```

Exemples for vector:

```
> myob@GN[1:20]
> slot(myob, "GN")[1:20]
> GN(myob)[1:20]
```

For data from the first array:

```
> myob[, 1]
```

2.4.3 Exclude data from object

The agExclude function allows to remplace by NA different data to exclude. There are different case:

- low-quality spots (thanks to the Flag slot of the object);
- controls if just the sample want to be observed;
- list of gene name to exclude (this list can be obtained by the argument identify=TRUE of the agMAplot or agImage functions which return a list of the identify gene names).

```
> M <- agExclude(myob, type = "controls", toNA = TRUE)
> M2 <- agExclude(myob, type = "flags", toNA = TRUE)
> a <- c("DarkCorner", "GE_BrightCorner")
> M3 <- agExclude(myob, type = "list", list = a, toNA = TRUE)
```

All these commands return an object of the same class whose the controls, flags or data of the gene name list are remplaced by NA.

3 Diagnostics plots

Several functions allow to visualise data of arrays:

- agBoxplot (boxplot of a Slot distribution)
- agMAplot (MA plot of a Slot)
- agImage (Virtual image of a Slot)
- agPlot (intensities values along the different chromosomes)

By default, if there is no whichSlot argument, the data used are:

- gM slot for an AgilentBatch object
- SgNorm slot for an AgilentNorm object
- logRatio for an AgilentBatchRG object which is calculating as log2(rM/gM)
- logRatioNorm for an AgilentNormRG object which is calculating as log2(rSgNorm/gSgNorm)

Moreover, it's possible to save these plots by the arguments p̈dfänd ḧtmläs TRUE. These plots are saved on the working directory.

3.1 The agBoxplot function

This function allow to observe the distribution of a slot. The result is different if they are one or several array. The argument of this function are:

```
> args(agBoxplot)

function (x, whichSlot = NULL, array = NULL, log = TRUE, centered = FALSE,
    reduced = FALSE, html = FALSE, pdf = FALSE, horizontal = FALSE)
NULL
```

3.1.1 Distribution inter-arrays

e.g. for several arrays. One boxplot is obtained for each array.

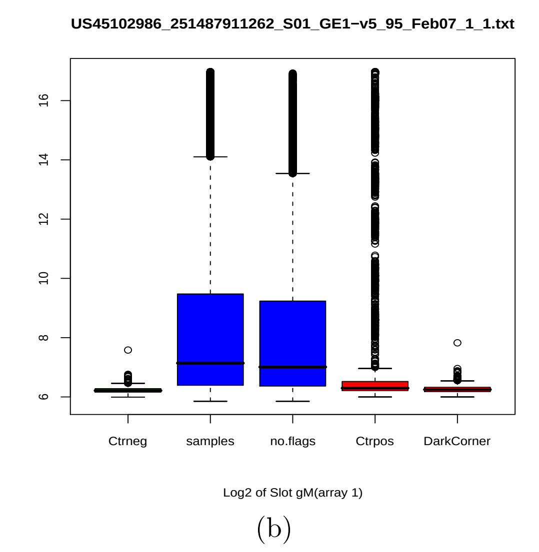
```
> agBoxplot(myob, array = 1:4)
```

3.1.2 Distribution intra-array

e.g. for just one array. Four boxplot are obtained to allow to observe the distribution of the negative controls, samples, positive controls and DarkCorner (which is the most important positive control on the array).

```
> agBoxplot(myob, array = 1)
```

(a)



(b)

Figure 1: The agBoxplot function. (A) Boxplot obtained for the gMeanSignal slot by default and for all the array of the AgilentBatch object. (B) Visualization of the negatives and positives controls distribution for the first array of the AgilentBatch object.

But the distributions can be obtain thanks to the summary function by the command:

```
> summary(myob)
```

```
Summary of:
        - slot gP:

       V1                   V2                   V3
 Min.    :4.487e+00   Min.    :4.535e+00   Min.    :3.415e+00
 1st Qu.:6.402e+00   1st Qu.:6.802e+00   1st Qu.:5.763e+00
 Median :5.353e+01   Median :5.704e+01   Median :6.347e+01
 Mean    :1.997e+03   Mean    :2.170e+03   Mean    :2.138e+03
 3rd Qu.:5.759e+02   3rd Qu.:6.300e+02   3rd Qu.:6.970e+02
 Max.    :1.594e+05   Max.    :1.697e+05   Max.    :1.521e+05
       V4
 Min.    :3.833e+00
 1st Qu.:5.578e+00
 Median :5.169e+01
 Mean    :1.980e+03
 3rd Qu.:5.575e+02
 Max.    :1.573e+05


        - slot gM:
```

```
      V1                  V2                 V3                  V4
 Min.   :    61.09   Min.   :    57.8   Min.   :    63.43   Min.   :    59.02
 1st Qu.:    81.17   1st Qu.:    82.9   1st Qu.:    83.61   1st Qu.:    83.56
 Median :   131.14   Median :   135.5   Median :   141.11   Median :   130.41
 Mean   :  2076.09   Mean   :  2252.2   Mean   :  2217.22   Mean   :  2063.26
 3rd Qu.:   652.93   3rd Qu.:   706.1   3rd Qu.:   772.31   3rd Qu.:   632.20
 Max.   :128480.40   Max.   :128453.9   Max.   :128744.30   Max.   :128945.60
```

```
       - slot gBGM:

      V1                V2                V3                  V4
 Min.   : 39.17   Min.   : 36.41   Min.   :   40.54   Min.   : 38.65
 1st Qu.: 46.12   1st Qu.: 46.32   1st Qu.:   46.50   1st Qu.: 46.19
 Median : 48.00   Median : 48.21   Median :   48.33   Median : 47.69
 Mean   : 49.07   Mean   : 49.32   Mean   :   52.24   Mean   : 48.29
 3rd Qu.: 50.48   3rd Qu.: 50.75   3rd Qu.:   50.65   3rd Qu.: 49.55
 Max.   :180.31   Max.   :226.48   Max.   :25870.13   Max.   :152.98
```

```
       - slot CtrT:


   -1     0     1
  153 43379  1486
```

```
       - slot Flag:

             name.array    ok   a c e  g i   j    l total.flags
Array 1 251487911262_1_1 42357 115 5 0 16 0 834 2570        2661
Array 2 251487911262_1_2 42421 128 3 0 28 0 872 2489        2597
Array 3 251487911262_1_3 42244 113 3 0 45 0 774 2661        2774
Array 4 251487911262_1_4 42921 113 2 0 31 0 328 2012        2097
```

```
Legend:

a        Feature is saturated                    glsSaturated

c        Feature is not uniform                  glsFeatNonUnifOL

e        Feature is not positive and significant  glsPosAndSignif

g        Feature is a population outlier         glsFeatPopnOL
```

```
i         Feature is manually marked              IsManualFlag

j         Background is not uniform               glsBGNonUnifOL

l         Background reading is population outlier glsBGPopnOL
```

And it's also possible to centered and reduced the inter-arrays data by the command:

```
> agBoxplot(myob, whichSlot = "gM", array = 1:4, centered = TRUE,
+     reduced = TRUE)
```

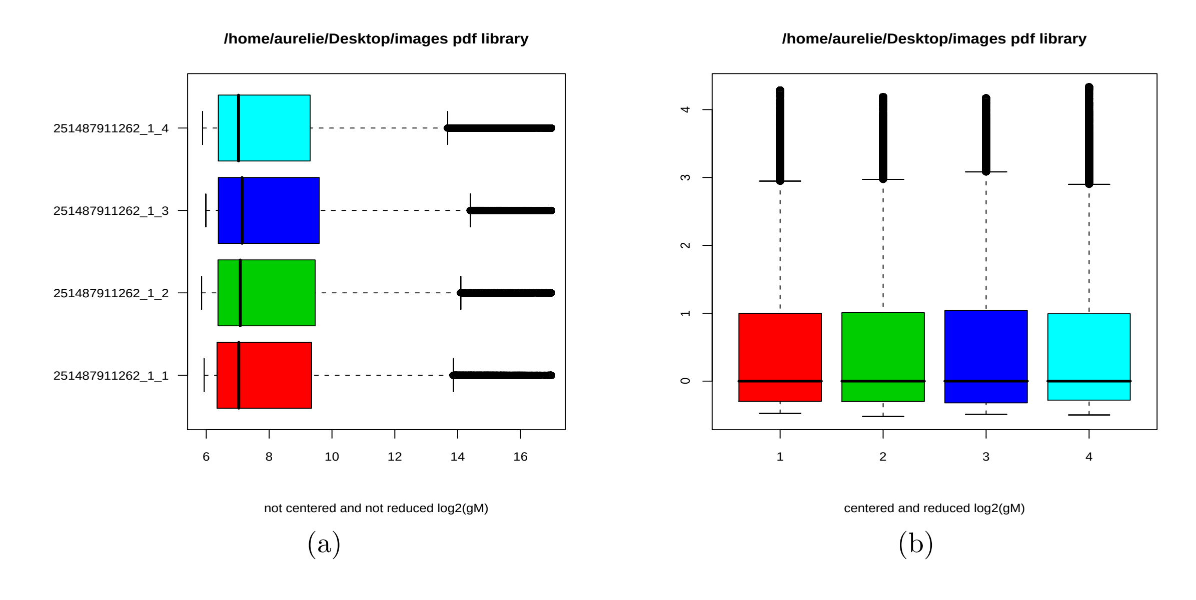


Figure 2: (A) Horizontal visualization of the agBoxplot for all the arrays. (B) Centered and reduced data from the first array of the AgilenBatch object.

3.2 The agMAplot function

This fonction allows to obtain a MA plot (Blant-Atlman plot) of the object for an array. A and M values are different in according to the class object. Indeed, for an AgilentBatch object, the reference correspond to a median of each spots in according to the different arrays, whereas for an AgilentBatchRG object, A and M are determined with the logRatio (log2(R/G)).
The different arguments of this function are:

```
> args(agMAplot)
```

A basic MA plot is obtained by the command:

```
> agMAplot(myob, whichSlot = "gM", array = 1)
```

Controls and Flags can be added to the MA plot by the command:

```
> agMAplot(myob, whichSlot = "gM", array = 1, ctr = TRUE, flag = 1:5)
```

Controls and distribution of A and M can be added to the MA plot by the command:

```
> agMAplot(myob, whichSlot = "gM", array = 1, ctr = TRUE, hist = TRUE)
```

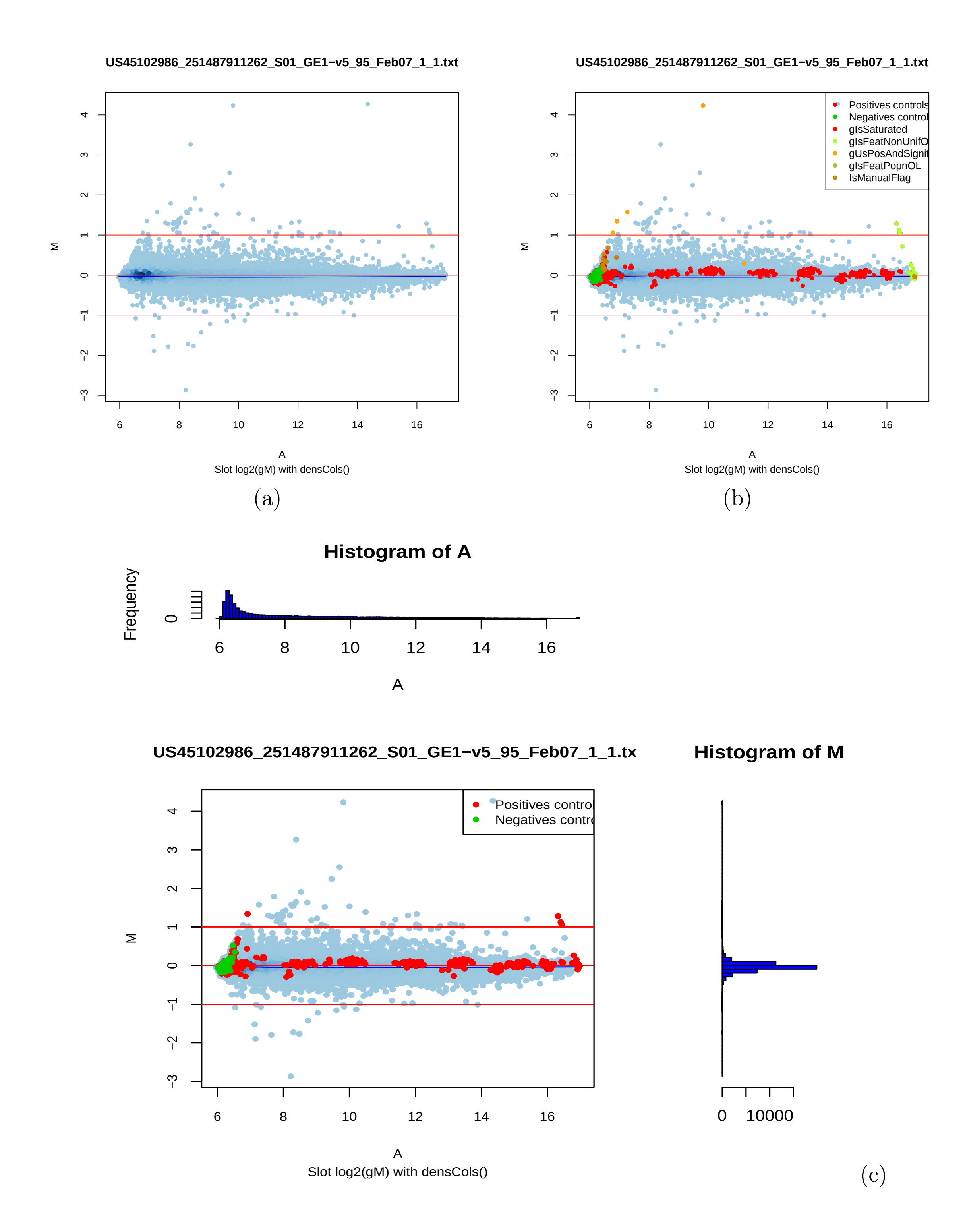


Figure 3: The agMAplot function. (A) MA plot obtain with the gMeanSignal slot and for the first array. (B) Visualization of the different flags on the MA plot in same time that the controls.(C) Visualization of controls and distribution of A and M on the MA plot (a).

The argument show.gene allows to observe the gene name of all the probe whose the M values aren't between, for this example, -1 et 1 (e.g. for the M value inferior to -1 and superior to 1). The command is:

```
> agMAplot(myob, whichSlot = "gM", array = 1, show.gene = c(-1,
+     1))
> a <- c("DarkCorner", "GE_BrightCorner")
> agMAplot(myob, whichSlot = "gM", array = 1, show.gene = a)
```

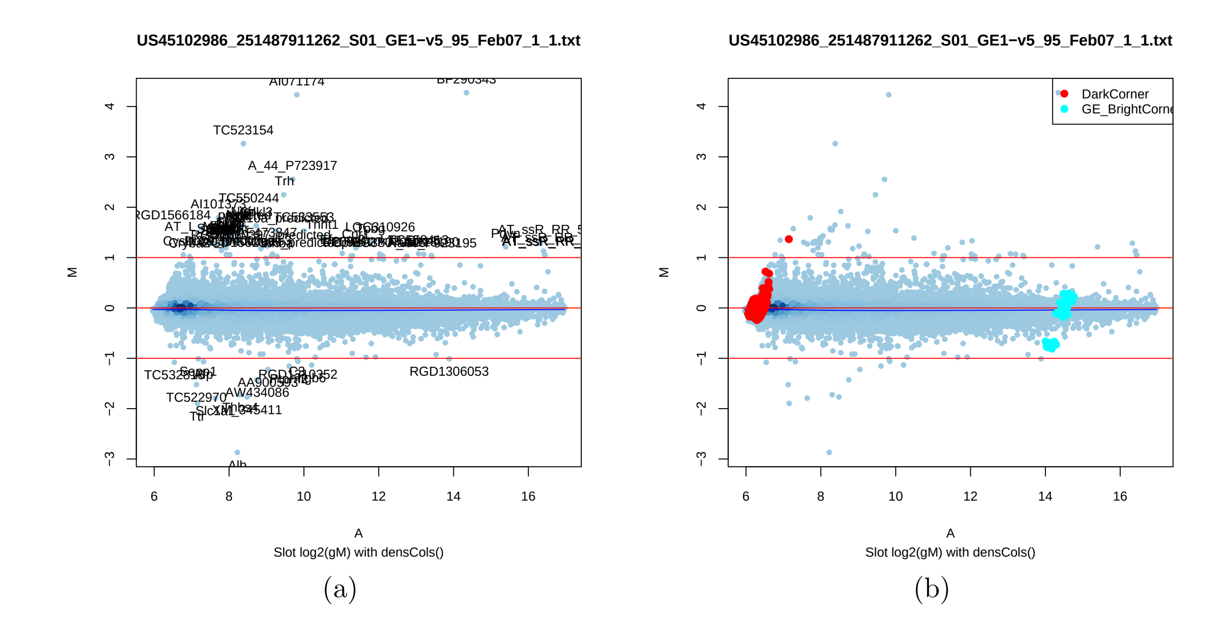


Figure 4: The agMAplot function. (A) Visualization of gene name (thanks to GeneName slot) for the M value inferior to -1 and superior to 1.(B) Visualization of the position on the plot of the spots corresponding to the list "a" of Gene Name.

3.3 The agImage function

This function allows to obtain a virtual image of each array. Thus, presence of paths of colour or anomaly can be observed. The different arguments of this function are:

```
> args(agImage)

function (object, whichSlot = NULL, array = 1, type = "intensity",
    threshold = NULL, bar = TRUE, html = FALSE, pdf = FALSE,
    identify = FALSE, col.zoom = NULL, row.zoom = NULL, log = TRUE,
    show.gene = NULL)
NULL
```

For a basic image, use this command:

```
> agImage(myob, whichSlot = "gM", array = 1)
```

Figure 5: Virtual image for the first array of the AgilentBatch object obtain thanks to the agImage function.

Another function (agThreshold function) allows to observe this image with a threshold. Thus all the intensity values under the threshold are remplace by NA (and are not observed on the image). This threshold can be defined in accordance with two methods:

- a percentage of intensity between minimum and maximum intensities values;
- intensities distribution e.g. in accordance with the quantiles (25, 50 ou 75 of the intensities values).

This function return the same object but with NA for the values inferior to threshold. This function can be directly used on the agImage function by the type and threshold arguments, by the command:

```
> agImage(myob, "gM", array = 1, type = "quartiles", threshold = 25)
> agImage(myob, "gM", array = 1, type = "intensity", threshold = 10)
```

As was mentioned previously, it's possible to exclude spots (controls, flags or list of gene names) by the agEclude function. And with this new object we can observe a new virtual image.
Moreover, an image, where a list of gene can be located, can be obtained by the command:

```
> a <- c("DarkCorner", "GE_BrightCorner")
> agImage(myob, "gM", array = 1, show.gene = a)
```

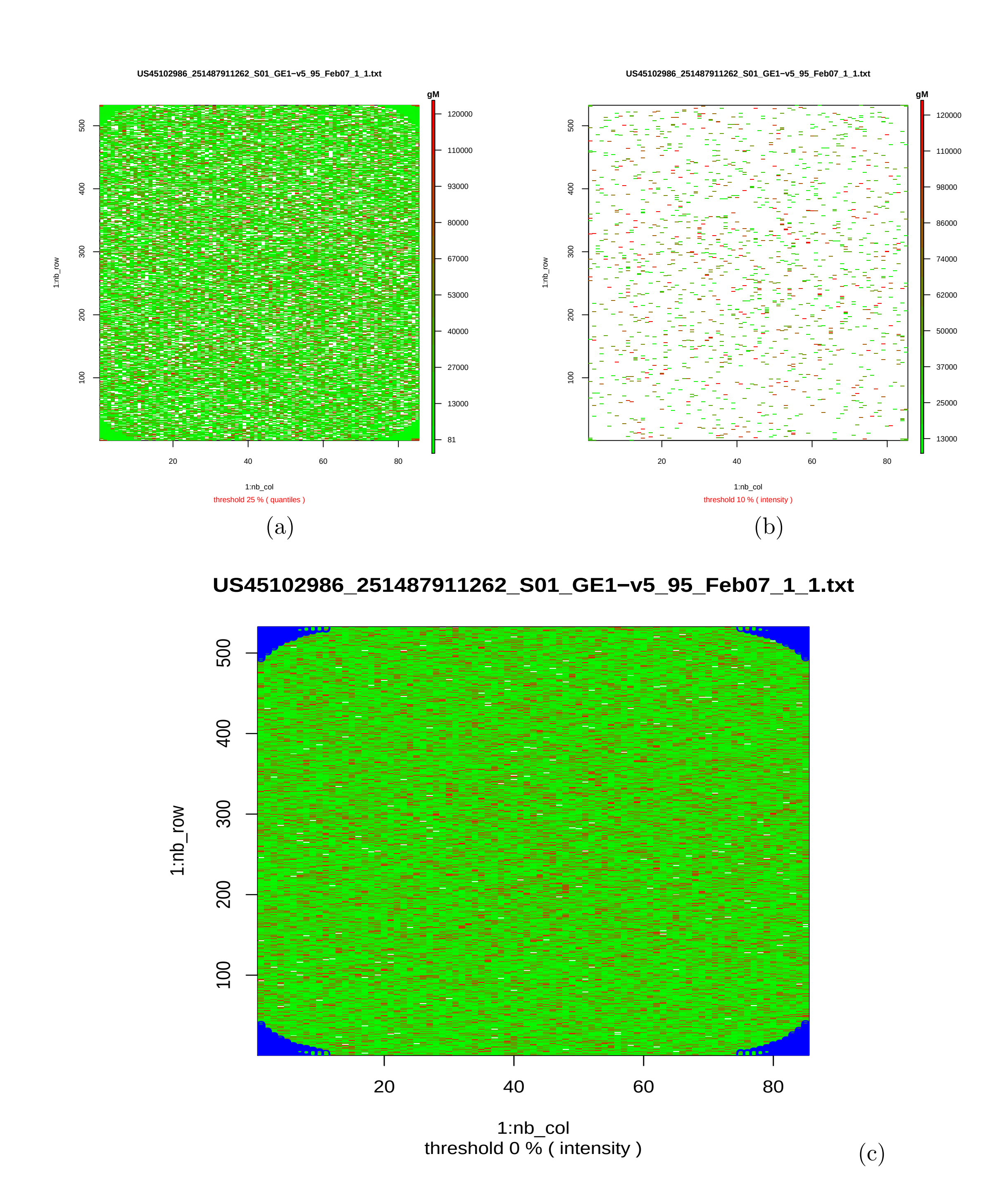


Figure 6: The others arguments of the agImage function. (A) Visualization of the virtual image where only the intensity upper than the first quantile e.g. 25 percent quantile are present. (B) Visualization of the same image but with a threshold of 10 percent in according with signal intensity. (C) Localisation of spots on the image corresponding to a list of gene name in blue circle, here DarkCorner and GE_BrightCorner which are two controls .

3.4 The agPlot function

This function is more specific of the two-colours array. Because, the plot obtained allows to observe a variation of the signal (e.g. logRatio for an AgilentBatchRG object or gM for an AgilentBatch) along a chromosome or on a sorted SN slot. The arguments of this function are:

```
> args(agPlot)

function (object, array = 1, whichSlot = NULL, log = TRUE, chrm = NULL,
    scale = NULL, barplot = TRUE, identify = FALSE, html = FALSE,
    pdf = FALSE)
NULL
```

For exemple for the 50th probes of the first chromosome, the plot obtained is:

```
> agPlot(myobRG, barplot = TRUE, log = FALSE, chrm = 1, scale = 1:50)
```

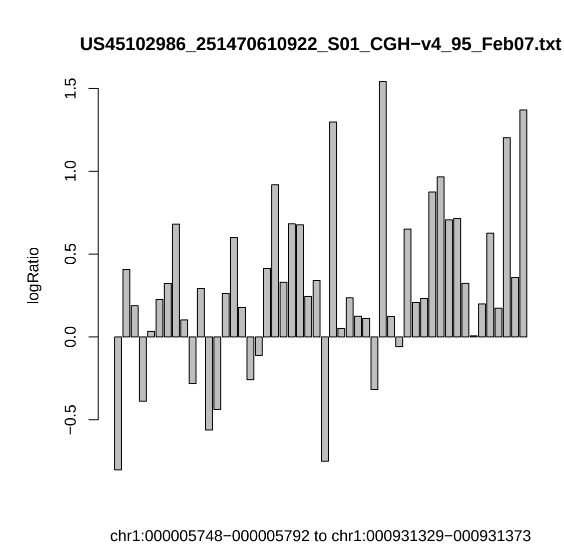


Figure 7: The agPlot function for the 50th probes of the first chromosome.

4 Normalization

Two methods of normalisation are used in this library:

- Lowess methods
- Quantiles methods

With these two methodes, a correction of the background signal is possible to improve the data normalization. This correction correspond to the argument bgCorrection of the agNormData function. The differents normalized slot obtained are in logarithm base 2. Moreover, two others slots are created: A and M which correspond to the values of A and M when the MA plot is plotted. Futhermore, a difference exists between normalization of one or two-colours array. Indeed, for a one-colour the reference correspond to the medians of each spots between the different arrays of the AgilentBatch or AgilentBatchRG object, whereas for a two-colors the reference of the red channel corespond to the green channel.

```
> args(agNormData)

function (object, whichSlot = c("Mean", "Processed"), bgCorrection = F,
    type = "quantiles", percent = 1)
NULL
```

4.1 Lowess method

This normalization can be obtained by the agLowessNorm function which retun a normalized object by the Lowess (or Loess) method without background correction. The command to obtain this normalized object are:

```
> norm1 <- agNormData(myob, "Mean", bgCorrection = FALSE, type = "lowess",
+     percent = 1)
> norm1 <- agLowessNorm(myob, "Mean")
```

To process the background correction, the command is:

```
> norm2 <- agNormData(myob, "Mean", bgCorrection = TRUE, type = "lowess",
+     percent = 1)
```

4.2 Quantiles method

This normalization can be obtained by the agQuantilesNorm function which retun a normalized object by the quantiles method without background correction. The command to obtain this normalized object (with or without background correction) are:

```
> norm3 <- agNormData(myob, "Mean", bgCorrection = FALSE, type = "quantiles")
> norm3 <- agQuantilesNorm(myob, "Mean")
> norm4 <- agNormData(myob, "Mean", bgCorrection = TRUE, type = "quantiles")
```

Thus, the normalized object class is:

- AgilentNorm for a normalized object from an AgilentBatch object class (one-colour);
- AgilentNormRG for a normalized object from an AgilentBatchRG object class (two-colours).

These normalized objects contain new slots: SgNorm, rSgNorm, gSgNorm, A, M and Type.

Slot name	Description
SgNorm	Normalised signal from the mean signal mesured on the green channel (gM slot) for an AgilentNorm object.
A	Matrix of the A values for a normalized object (AgilentNorm or AgilentormRG)
M	Matrix of the M values for a normalized object (AgilentNorm or AgilentormRG)
Type	Method of normalization used to obtain the normalized object (AgilentNorm or AgilentormRG)

Table 3: Table of the new specific slots of the normalized object (AgilentNorm and AgilentNormRG).

After normalization, agMAplot can be use to observe the effect of this normalization on data. And the new data distribution can be observed by agBoxplot function.
Finally, a new virtual image can be visualized.

5 Data exportation

This exportation of data is realized by the transformation of the object (normally the normalized object) in a ExpressionSet object, since the R-2.5.0 version. Thus this new object allows to use the other library of Bioconductor (for example Biobase)

5.1 Creation of the ExpressionSet object

Before to perform an ExpressionSet conversion, the flag matrix must be change in a booleen matrix. If only the second type of flags is necessary, users can use the command:

```
> myob <- agConvFlag(myob, flag = 2)
```

The Transformation of an AgilentBatch, AgilentNorm, AgilentBatchRG or AgilentNormRG objects in an ExpressionSet object is processed by the following command:

```
> es <- as(myob, "ExpressionSet")
```

```
This ExpressionSet object from an AgilentBatch object (one-colour arrays) is saved in:
         /home/aurelie/R/i686-pc-linux-gnu-library/2.5/AgiData/ExpressionSet.txt
Flags corresponding to the data from an AgilentBatch object (one-colour arrays) are saved in:
         /home/aurelie/R/i686-pc-linux-gnu-library/2.5/AgiData/Flags.txt
```

5.2 Exemples using the ExpressionSet object

Then other library can be used, like Biobase:

- Display gene name:

```
> featureNames(es)[1:25]
```

```
 [1] "GE_BrightCorner"      "DarkCorner"           "DarkCorner"
 [4] "DarkCorner"           "DarkCorner"           "DarkCorner"
 [7] "DarkCorner"           "DarkCorner"           "DarkCorner"
[10] "DarkCorner"           "DarkCorner"           "AA892298"
[13] "AI232741"             "Gmpr"                 "XM_236342"
[16] "RGD1309888"           "XM_222163"            "AA891661"
[19] "Plcb3"                "Polr2b_predicted"     "XM_225162"
[22] "F10"                  "Btbd5_predicted"      "RGD1310717_predicted"
[25] "Prkr"
```

- Visualization of expression data:

```
> exprs(es)[1:25]
```

```
 [1] 20354.110000    11.306570     5.485073     5.550568     5.610573
 [6]     5.665902     7.786766     8.300380     5.807561     5.845899
[11]     5.880494    10.634210    10.221720  2621.380000    11.501160
[16]  1173.688000  2910.492000    16.168080   203.756200   690.069400
[21]     6.061691    18.307080    12.981660     6.073625    22.823440
```

ANNEXE B

Manuel d'utilisation de la librairie R/Bioconductor : RTools4TB

The RTools4TB package: data mining of public microarray data through connections to the TranscriptomeBrowser database.

A. Bergon, F. Lopez, J. Textoris, S. Granjeaud and D. Puthier

October 28, 2009

TAGC/Inserm U928. Parc Scientifique de Luminy case 928. 163, avenue de Luminy. 13288 MARSEILLE cedex 09. FRANCE
bergon@tagc.univ-mrs.fr

Contents

1 Overview

TranscriptomeBrowser (TBrowser, `http://tagc.univ-mrs.fr/tbrowser`) hosts a large collection of transcriptional signatures (TS) automatically extracted from the Gene Expression Omnibus (GEO) database. Each GEO experiment (GSE) was processed so that a subset of the original expression matrix containing the most relevant/informative genes was kept and organized into a set of homogeneous signatures [1]. Each signature was tested for functional enrichment using annotations terms obtained from numerous ontologies or curated databases (Gene Ontology, KEGG, BioCarta, Swiss-Prot, BBID, SMART, NIH Genetic Association DB, COG/KOG...) using the DAVID knowledgebase [2].
RTools4TB is a library for data mining of public microarray data. *RTools4TB* can be helpful (i) to define the biological contexts (i.e, experiments) in which a set of genes are co-expressed and (ii) to define their most frequent neighbors [1]. The RTools4TB package also implements the DBF-MCL algorithm ("Density Based Filtering And Markov Clustering") that can be used for fast and automated partitioning of microarray data. DBF-MCL is a tree-step adaptative algorithm that (i) find elements located in dense

areas (ie. clusters) (ii) uses selected items to construct a graph and (iii) performs graph partitioning using MCL [3]. Note that a UNIX-like systems is required to use DBF-MCL.

2 Fetching transcriptional signatures from TBrowserDB

2.1 The getSignatures function

Connection to the TranscriptomeBrowser database (TBrowserDB) relies on the `getSignatures`, `getExpressionMatrix` and `getTBInfo` functions.

Basically, the `getSignatures` function can be used to retrieve transcriptional signature IDs using gene symbol(s), probe ID(s), experiment ID, microarray platform ID or annotation term(s) as input. This is controled by the "field" argument.

```
> library(RTools4TB)
> args(getSignatures)

function (field = c("gene", "probe", "platform", "experiment",
    "annotation"), value = NULL, qValue = NULL, nbMin = NULL,
    verbose = TRUE, save = FALSE)
NULL
```

Once the field argument is set, one need to provide a value as input. For instance the following query use gene name as input with value "PCNA".

```
> res <- getSignatures(field = "gene", value = "PCNA")
> head(res)
```

Transcriptional signature IDs can also be obtained by selecting the relevant experiment IDs, platform IDs and probe IDs. To get all transcriptional signature IDs associated with GSE2004 experiment, one should use the following syntax:

```
> res <- getSignatures(field = "experiment", value = "GSE2004")

23 signatures were found for the request:
 GSE  =  GSE2004
```

To get all signatures obtained on GPL96 platform, use the following syntax:

```
> res <- getSignatures(field = "platform", value = "GPL96")

3377 signatures were found for the request:
 GPL  =  GPL96
```

Moreover, as all signatures were tested for functional enrichment using keywords from the DAVID knowledgebase these terms can be used to query the database (DAVID collects a wide range of annotation from several databases including: GO, BIOCARTA, KEGG, PANTHER, BBID,...) . The `annotationList` dataset contains the annotations terms.

```
> data(annotationList)
> names(annotationList)

[1] "Keyword"   "TableName"
```

```
> attach(annotationList)
> annotationList[1:4, ]
```

```
                       Keyword TableName
1               1.RBphosphoE2F      BBID
2  100.MAPK_signaling_cascades      BBID
3        104.Insulin_signaling      BBID
4 105.Signaling_glucose_uptake      BBID
```

```
> table(TableName)
```

```
TableName
                 BBID                BIOCARTA        COG_KOG_ONTOLOGY
                   57                     468                      22
             CYTOBAND GENETIC_ASSOCIATION_DB           GOTERM_BP_ALL
                  526                      68                    1273
        GOTERM_CC_ALL           GOTERM_MF_ALL           INTERPRO_NAME
                  328                     639                     777
         KEGG_PATHWAY           KEGG_REACTION           OMIM_PHENOTYPE
                  334                      78                      10
      PANTHER_PATHWAY               PFAM_NAME     PIR_HOMOLOGY_DOMAIN
                  104                     531                      86
  PIR_SUPERFAMILY_NAME               PUBMED_ID              SMART_NAME
                  160                    4887                     235
      SP_PIR_KEYWORDS
                  568
```

The selected terms can be used to select TS IDs. In this case, user should define a q-value. For instance one can select TS enriched in genes related to the "HSA04110:CELL CYCLE" KEGG pathway with q-value below $10e^{-20}$.

```
> cc <- getSignatures(field = "annotation", value = "HSA04110:CELL CYCLE",
+     qValue = 20)
```

```
66 signatures were found for the request:
 annotation  =  HSA04110:CELL CYCLE
```

Of note, one can also search for TS IDs containing genes located in the same chromosomal region. For instance one can select TS IDs enriched in genes located in the "8q" region which is frequently amplified or deleted in tumors. This will point out the biological contexts in which sets of genes located in the 8q region share the same expression profile, suggesting amplifications or deletions in some biological samples.

```
> query <- paste(grep("^8q", Keyword, val = T), collapse = "|")
> query
```

```
[1] "8q13|8q21|8q21.11|8q21.2|8q22.1|8q22.3|8q24|8q24.13|8q24.3"
```

```
> cc <- getSignatures(field = "annotation", value = query, qValue = 10)
```

```
4 signatures were found for the request:
 annotation  =  8q13|8q21|8q21.11|8q21.2|8q22.1|8q22.3|8q24|8q24.13|8q24.3
```

Next sections will introduce more complexe queries using sets of genes with or without Boolean operators.

2.1.1 Request without logical operators (gene list)

When field is set to "gene" or "probe", user can perform a request using a list of item separated by blanks. These blanks are interpreted as the OR logical operators. In this case, all signatures containing at least one gene of the list will be returned. To select more informative signatures we suggest to use the nbMin argument that will select signatures containing at least nbMin genes out of the list.

The following examples search for signatures containing at least 2 genes of the input list (CD3D, CD3E and CD4).

```
> gl <- getSignatures(field = "gene", value = "CD3D CD3E CD4",
+     nbMin = 2)
```

```
150 signatures were found for the request:
 gene  =  CD3D CD3E CD4 and nbMin =  2
```

```
> head(gl)
```

```
  Signature nb.Genes
1 03AD63FB5        2
2 050367D10        2
3 053ECFACF        3
4 05F2203B7        2
5 0C0A8F888        2
6 0D2EA9D52        2
```

2.1.2 Request using logical operators

The "value" argument of `getSignatures` may contain the following Boolean operators (see help section on TranscriptomeBrowser web site for more informations, `http://tagc.univ-mrs.fr/tbrowser`)

- & : AND
- | : OR
- ! : NOT , (used in conjonction with &)

This is a convenient way to create relevant queries. Suppose your field of interest is related to T-cell activation. You could be interested in retrieving all TS IDs that contain the CD4 gene as they likely contain additional T cell markers. Comparing these TS IDs should help you to define frequent CD4 neighbors (very likely related to TCR signaling cascade). Thereby, your request should be:

```
> res <- getSignatures(field = "gene", value = "CD4")
```

```
371 signatures were found for the request:
 gene  =  CD4
```

This gene is found in 371 TS (with the current database release). Obtaining associated gene lists would be time consuming and would not be as specific as expected. Indeed, the CD4 marker is also expressed by macrophages. Another solution would be to search for TS containing two T-cell markers (CD4 and CD3E for instance) and to exclude (using the NOT operator) those containing the CD14 marker (a macrophages marker). The syntax should be the following:

```
> res <- getSignatures(field = "gene", value = "CD4 & CD3E & !CD14")
```

```
55 signatures were found for the request:
 gene  =  CD4 & CD3E & !CD14
```

In the same way, one can try to exclude TS containing B-cells markers by discarding those containing the CD19 or IGHM markers. The resulting query would be the following:

```
> res <- getSignatures(field = "gene", value = "CD4 & CD3E & !(CD19 | IGHM)")

33 signatures were found for the request:
 gene  =  CD4 & CD3E & !(CD19 | IGHM)
```

2.2 Finding the biological contexts in which sets of genes are co-expressed

As mentioned by Lacroix *et al.*, ESR1, GATA3, XBP1 are co-expressed in breast cancer tumors (see [4]). This assumption can be easily verified using *RTools4TB*. For instance, in the following examples, we fetch transcriptional signature IDs that contain "XBP1 & ESR1 & GATA3". Next the `getTBInfo` function is used to retrieve the experiment description from which they are derived (here only for TS ID "3DE64836D").

```
> TS <- getSignatures(field = "gene", value = "XBP1 & ESR1 & GATA3")

14 signatures were found for the request:
 gene  =  XBP1 & ESR1 & GATA3

> head(TS)

[1] "0F2635383" "3DE64836D" "59A18E225" "8059848B4" "84E5E1077" "8F69864F9"

> a <- getTBInfo(field = "signature", value = "3DE64836D", verbose = FALSE)
> exp <- a["Experiment", 1]
> info <- getTBInfo(field = "experiment", value = exp, verbose = TRUE)

A result was found for : experiment  =  GSE7904
Name = GSE7904
Organism = Homo sapiens
PMID = NULL
Nb. samples = 62
Title = Expression data from human breast tissue
Summary = bulk breast tumor RNA from patientAbstract: Sporadic basal-like
     cancers (BLC) are a distinct class of human breast cancers that are
     phenotypically similar to BRCA1-associated cancers. Like BRCA1-deficient
     tumors, most BLC lack markers of a normal inactive X chromosome (Xi).
     Duplication of the active X chromosome and loss of Xi characterized almost
     half of BLC cases tested. Others contained biparental but
     nonheterochromatinized X chromosomes or gains of X chromosomal DNA. These
     abnormalities did not lead to a global increase in X chromosome
     transcription but were associated with overexpression of a small subset of
     X chromosomal genes. Other, equally aneuploid, but non-BLC rarely
     displayed these X chromosome abnormalities. These results suggest that X
     chromosome abnormalities contribute to the pathogenesis of BLC, both
     inherited and sporadic.total 62 sample incudes 43 tumor, 7 normal breast
     and 12 normal organelle
```

As expected the transcriptional signature "3DE64836D" correspond to a breast cancer tumor analysis. This is also true for the other TS (not shown).

2.3 Finding transcriptional neighbors

One interesting feature of *RTools4TB* its ability to find genes frequently co-expressed with the input list. Indeed, results from a request to TBrowserDB can be displayed as a graph using the `createGraph4BioC` function. This function retrieves the list of TS that verify the constrain (here "XBP1 & ESR1 & GATA3"). A list of gene falling in at least one of the TS is next computed. A gene-gene matrix M is created that will record for each pair of gene the number of time they were observed in the same signature. In the following example, only a subset of this adjacency matrix (containing genes falling in a significant proportion of signatures, `prop=80%`) is used to create a graph.

```
> library(biocGraph)

> adjMat <- createGraph4BioC(request = "XBP1 & ESR1 & GATA3", prop = 80)

> g1 <- new("graphAM", adjMat = adjMat)
> nodes(g1)

 [1] "C6orf211"  "GREB1"     "WWP1"      "JMJD2B"    "KRT18"     "RNF103"
 [7] "ROGDI"     "SLC22A5"   "THSD4"     "NAT1"      "SLC39A6"   "ABAT"
[13] "CA12"      "CIRBP"     "LOC400451" "MAGED2"    "MCCC2"     "MLPH"
[19] "ANXA9"     "ERBB4"     "FOXA1"     "ESR1"      "GATA3"     "TBC1D9"
[25] "XBP1"

> nAt <- makeNodeAttrs(g1)
> nAt$fillcolor[match(rownames(as.matrix(nAt$fillcolor)), c("GATA3",
+     "XBP1", "ESR1"), nomatch = F) != 0] <- "green"
> nAt$fillcolor[match(rownames(as.matrix(nAt$fillcolor)), c("TBC1D9",
+     "FOXA1"), nomatch = F) != 0] <- "yellow"
> plot(g1, "fdp", nodeAttrs = nAt)
```

As expected the list of gene contains "XBP1 & ESR1 & GATA3" but also FOXA1/HNF3A that was reported to be co expressed with ESR1 in several experiments (see [4]). Other genes are also particularly relevant such as TBC1D9/MDR1 (Multidrug Resistance 1) (figure 1).

2.4 Vizualising expression matrix.

The TS "3DE64836D" is related to experiment "GSE7904". In this experiments, the authors were interested in analysing several classes of breast cancer tumors especially "Sporadic basal-like cancers".

```
> a <- getTBInfo(field = "signature", value = "3DE64836D", verbose = FALSE)
> exp <- a["Experiment", 1]
> info <- getTBInfo(field = "experiment", value = exp, verbose = TRUE)

A result was found for : experiment  =  GSE7904
Name = GSE7904
Organism = Homo sapiens
PMID = NULL
Nb. samples = 62
Title = Expression data from human breast tissue
Summary = bulk breast tumor RNA from patientAbstract: Sporadic basal-like
     cancers (BLC) are a distinct class of human breast cancers that are
     phenotypically similar to BRCA1-associated cancers. Like BRCA1-deficient
     tumors, most BLC lack markers of a normal inactive X chromosome (Xi).
```

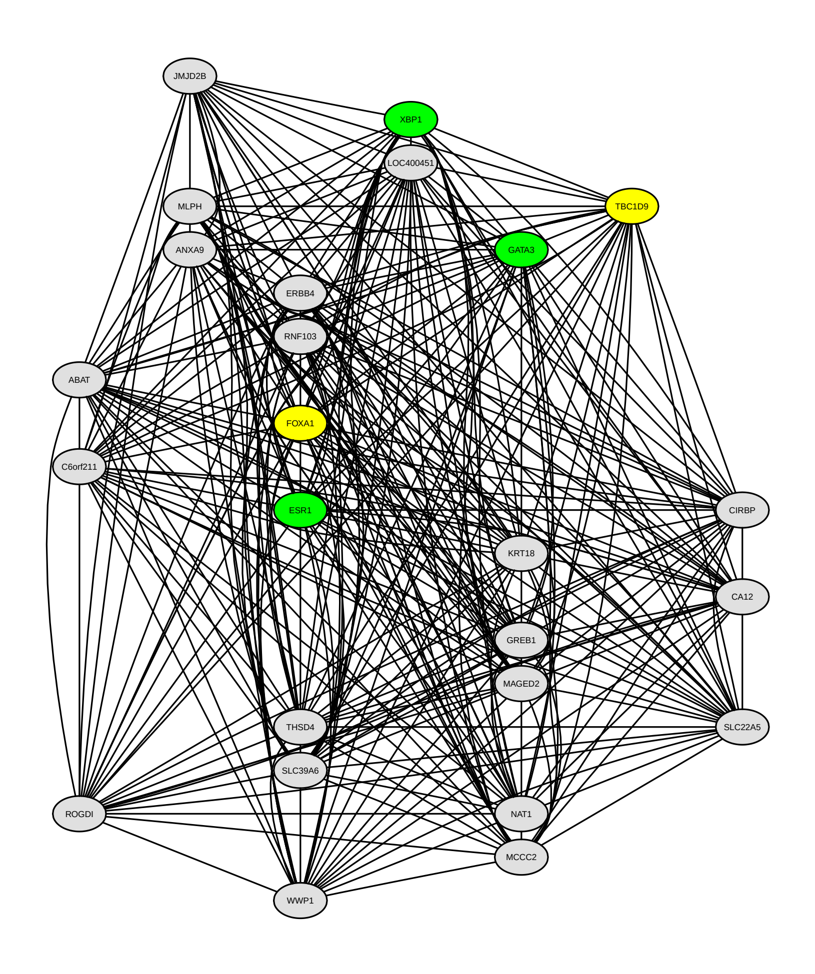


Figure 1: A graph containing "XBP1 & ESR1 & GATA3" together with their most frequent transcriptional neighbors.

```
Duplication of the active X chromosome and loss of Xi characterized almost
half of BLC cases tested. Others contained biparental but
nonheterochromatinized X chromosomes or gains of X chromosomal DNA. These
abnormalities did not lead to a global increase in X chromosome
transcription but were associated with overexpression of a small subset of
X chromosomal genes. Other, equally aneuploid, but non-BLC rarely
displayed these X chromosome abnormalities. These results suggest that X
chromosome abnormalities contribute to the pathogenesis of BLC, both
inherited and sporadic.total 62 sample incudes 43 tumor, 7 normal breast
and 12 normal organelle
```

The samples that were used are the following.

```
> sampleInfo <- getTBInfo(field = "samples", value = "3DE64836D")
> head(sampleInfo[, 1:2])

     sampleID    Title
[1,] "GSM194397" "Basal (T118)"
[2,] "GSM194398" "Basal (T134)"
[3,] "GSM194399" "Basal (T140)"
[4,] "GSM194400" "Basal (T141)"
[5,] "GSM194401" "Basal (T146)"
[6,] "GSM194402" "Basal (T147)"
```

Using the getExpressionMatrix function, the expression matrix for signature ”3DE64836D” can be fetched in order to visualize the expression profile of ESR1, GATA3 and XBP1 tumors compare to normal breast tissues.

```
> em <- getExpressionMatrix(signatureID = "3DE64836D")

Downloading expression matrix for transcriptional signature:  3DE64836D ( 62 samples x 143 probes)

> class(em)

[1] "data.frame"
```

The getExpressionMatrix function returns a data.frame. The first two columns store probe IDs and gene symbols. Additional columns contain corresponding expression values (figure 2).

```
> library(RColorBrewer)
> col <- colorRampPalette(brewer.pal(10, "RdBu"))(256)
> geneNames <- paste(em[, 1], em[, 2], sep = "||")
> em <- as.matrix(em[, -c(1, 2)])
> ind <- match(colnames(em), sampleInfo[, 1])
> colnames(em) <- sampleInfo[ind, 2]
> row <- rep(1, nrow(em))
> ind <- grep("(XBP1)|(ESR1)|(GATA3)", geneNames, perl = TRUE)
> row[ind] <- 2
> rc <- rainbow(2, start = 0, end = 0.3)
> rc <- rc[row]
> col <- colorRampPalette(brewer.pal(10, "RdBu"))(256)
> split <- strsplit(colnames(em), " (", fixed = TRUE)
> pheno <- unlist(lapply(split, "[", 1))
```

```
> pheno <- as.factor(pheno)
> levels(pheno) <- 1:5
> cc <- rainbow(5, start = 0, end = 0.3)
> cc <- cc[pheno]
> heatmap(em, col = col, RowSideColors = rc, ColSideColors = cc,
+     labRow = geneNames, cexRow = 0.3)
```

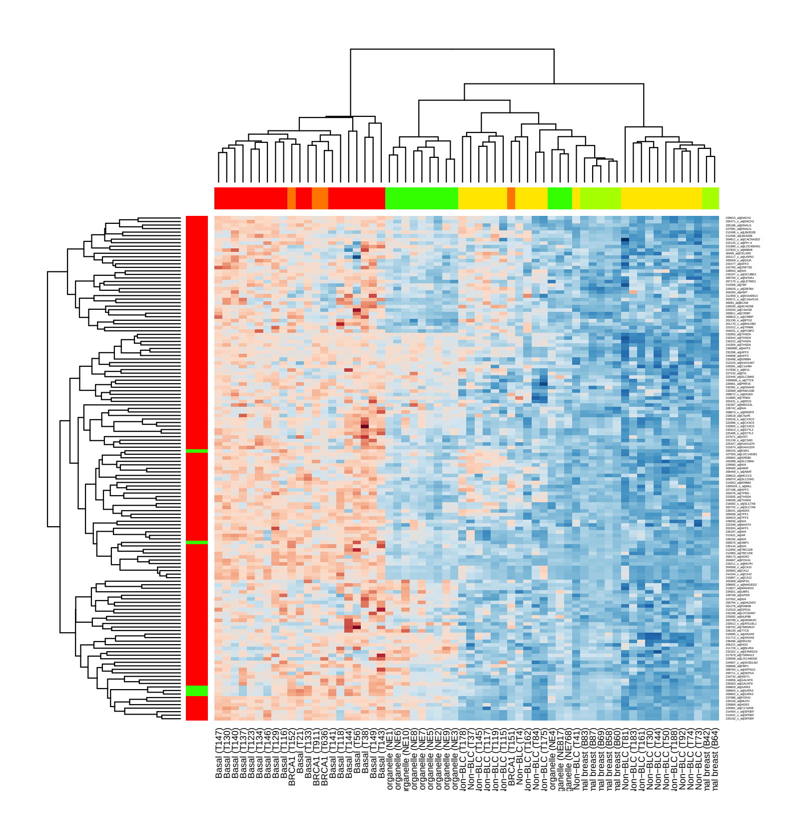


Figure 2: The expression matrix corresponding to signature "3DE64836D".

Of note, the `plotGeneExpProfiles` is a high level function to visualize gene expression levels in a signature (figure 3).

```
> plotGeneExpProfiles(data = em, X11 = FALSE)
```

3 Creating transcriptional signatures from a user defined data set using DBF-MCL algorithm.

When analyzing a noisy dataset, one is interested in isolating dense regions as they are populated with genes/elements that display weak distances to their nearest neighbors (i.e. strong profile similarities). To isolate these regions DBF-MCL computes, for each gene/element, the distance with its kth nearest neighbor (DKNN). In order to define a critical DKNN value that will depend on the dataset and below which a gene/element will be considered as falling in a dense area, DBF-MCL computes simulated DKNN values by using an empirical randomization procedure. Given a dataset containing n genes

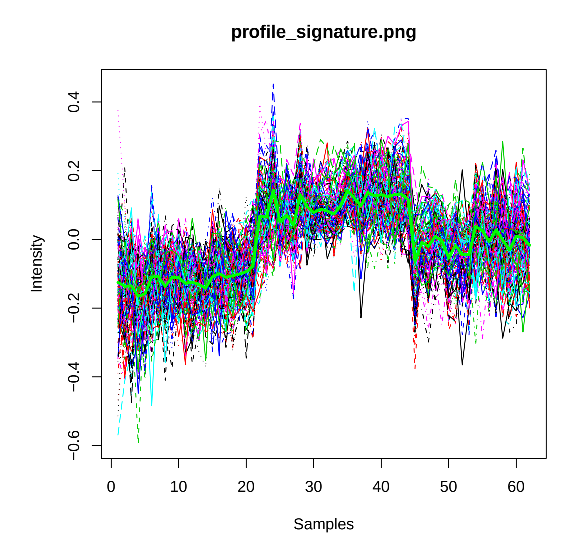


Figure 3: Gene expression profiles of signatures containing XBP1, ESR1 and GATA3: the centroid is highlighted in green.

and p samples, a simulated DKNN value is obtained by sampling n distance values from the gene-gene distance matrix D and by extracting the kth-smallest value. This procedure is repeated n times to obtain a set of simulated DKNN values S. Computed distributions of simulated DKNN are used to compute a FDR value for each observed DKNN value. The critical value of DKNN is the one for which a user-defined FDR value (typically 10%) is observed. Genes with DKNN value below this threshold are selected and used to construct a graph. In this graph, edges are constructed between two genes (nodes) if one of them belongs to the k-nearest neighbors of the other. Edges are weighted based on the respective coefficient of correlation (*i.e.*, similarity) and the graph obtained is partitioned using the Markov CLustering algorithm (MCL).

3.1 Installation

With the current implementation DBFMCL function works only on UNIX-like plateforms. MCL is required and can be installed using your package manager or using the following command lines pasted in a terminal:

```
# Download the latest version of mcl
# (the library has been tested successfully with the 06-058 version)
wget http://micans.org/mcl/src/mcl-latest.tar.gz
# Uncompress and install mcl
tar xvfz mcl-latest.tar.gz
cd mcl-xx-xxx
./configure
make
sudo make install
# You should get mcl in your path
mcl -h
```

3.2 Examples

We will search for transcriptional signatures in a subset of the ALL dataset.

```
> library(ALL)
> data(ALL)
> sub <- exprs(ALL)[1:3000, ]
```

First, we will normalize the data set using the `doNormalScore` function. This function performs normal score transformation of a matrix. The `doNormalScore` transforms each sample to follow a normal distribution (with `mean = 0` and `sd = 1`). Alternatively, users may also use other normalization routines such as : `doRankTransformation` or `limma::normalizeQuantiles`.

```
> subNorm <- doNormalScore(sub)
```

The `DBFMCL` function allows one to extract TS from a data set. Its behaviour is controlled by several arguments.

```
> args(DBFMCL)

function (data = NULL, filename = NULL, path = ".", name = NULL,
    distance.method = c("pearson", "spearman", "euclidean", "spm",
        "spgm"), clustering = TRUE, silent = FALSE, verbose = TRUE,
    k = 150, random = 3, memory.used = 1024, fdr = 10, inflation = 2,
    set.seed = 123, returnRank = FALSE)
NULL
```

The `DBFMCL` function accepts a tab-delimited file (argument `filename`), an expressionSet, a data.frame or a matrix (argument `data`) as input. The input data must contain an expression matrix with gene as rows and samples as columns. Note that space characters inside gene names are not allowed (as they are not supported by the mcl command-line program).

The two main parameters of DBF-MCL are `k` that controls the size of the neighborhood and the `inflation` (range 1.1 to 5) which controls the way the underlying graph is partitioned. In the following example, the neighborhood size (`k`) is set to 150 and the MCL inflation parameters is set to 2.0 (default MCL setting). Most generally these default parameters give very good results on microarray datasets. For a detailed discussion about these parameters please read the section ”Performances of DBF-MCL on Complex9RN200 dataset” in the article describing TranscriptomeBrowser stategy [1]. Morevover, in our example, the distance method is set to ”pearson” although the ”spearman” (that is the default method for computing TS in the TranscriptomeBrowser projet) also give very relevant results. Note that additional distance, including ”euclidean” and two mixtures of pearson” and ”spearman” (”spm” and ”spgm”) are also available.

```
> res <- DBFMCL(subNorm, distance.method = "pearson", memory = 512)
```

The results are stored in an instance of class `DBFMCLresult`.

```
> class(res)

[1] "DBFMCLresult"
attr(,"package")
[1] "RTools4TB"

> res
```

Bibliographie

[Hum, 2010] (2010) Human genome : Genomes by the thousand. *Nature,* **467** (7319), 1026–7.

[Aburatani, 2011] Aburatani, H. (2011) [Cancer genome analysis through next-generation sequencing]. *Gan to kagaku ryoho. Cancer & chemotherapy,* **38** (1), 1–6.

[Aerts *et al.*, 2006] Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.-C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P. & Moreau, Y. (2006) Gene prioritization through genomic data fusion. *Nature biotechnology,* **24** (5), 537–44.

[Ahmadian *et al.*, 2006] Ahmadian, A., Ehn, M. & Hober, S. (2006) Pyrosequencing : history, biochemistry and future. *Clinica chimica acta ; international journal of clinical chemistry,* **363** (1-2), 83–94.

[Al-Shahrour *et al.*, 2007] Al-Shahrour, F., Minguez, P., Tárraga, J., Medina, I., Alloza, E., Montaner, D. & Dopazo, J. (2007) FatiGO + : a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic acids research,* **35** (Web Server issue), W91–6.

[Alston *et al.*, 2010] Alston, M. J., Seers, J., Hinton, J. C. D. & Lucchini, S. (2010) BABAR : an R package to simplify the normalisation of common reference design microarray-based transcriptomic datasets. *BMC bioinformatics,* **11**, 73.

[Altschul *et al.*, 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) Basic local alignment search tool. *Journal of molecular biology,* **215** (3), 403–10.

[Ashburner *et al.*, 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000) Gene ontology : tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics,* **25** (1), 25–9.

[Bailey, 2011] Bailey, T. L. (2011) DREME : motif discovery in transcription factor ChIP-seq data. *Bioinformatics (Oxford, England),* **27** (12), 1653–9.

[Bainbridge *et al.*, 2011] Bainbridge, M. N., Wang, M., Wu, Y.-Q., Newsham, I., Muzny, D. M., Jefferies, J. L., Albert, T. J., Burgess, D. L. & Gibbs, R. A. (2011) Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome biology,* **12** (7), R68.

[Barnett *et al.*, 2011] Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. & Marth, G. T. (2011) BamTools : a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics (Oxford, England),* **27** (12), 1691–2.

[Baron *et al.*, 2011] Baron, D., Magot, A., Ramstein, G., Steenman, M., Fayet, G., Chevalier, C., Jourdon, P., Houlgatte, R., Savagner, F. & Pereon, Y. (2011) Immune response and mitochondrial metabolism are commonly deregulated in DMD and aging skeletal muscle. *PloS one,* **6** (11), e26952.

[Barozzi *et al.*, 2011] Barozzi, I., Termanini, A., Minucci, S. & Natoli, G. (2011) Fish the ChIPs : a pipeline for automated genomic annotation of ChIP-Seq data. *Biology direct,* **6** (1), 51.

[Barrett *et al.*, 2005] Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W. & Edgar, R. (2005) NCBI GEO : mining millions of expression profiles–database and tools. *Nucleic acids research,* **33** (Database issue), D562–6.

[Barski *et al.*, 2007] Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. & Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell,* **129** (4), 823–37.

[Bernstein *et al.*, 2010] Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S. & Thomson, J. A. (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology,* **28** (10), 1045–8.

[Bertos & Park, 2011] Bertos, N. R. & Park, M. (2011) Breast cancer - one term, many entities ? *The Journal of clinical investigation,* **121** (10), 3789–96.

[Bertucci *et al.*, 2004] Bertucci, F., Finetti, P., Rougemont, J., Charafe-Jauffret, E., Nasser, V., Loriod, B., Camerlo, J., Tagett, R., Tarpin, C., Houvenaeghel, G., Nguyen, C., Maraninchi, D., Jacquemier, J., Houlgatte, R., Birnbaum, D. & Viens, P. (2004) Gene expression profiling for molecular characterization of inflammatory breast cancer and prediction of response to chemotherapy. *Cancer research,* **64** (23), 8558–65.

[Bhinge *et al.*, 2007] Bhinge, A. A., Kim, J., Euskirchen, G. M., Snyder, M. & Iyer, V. R. (2007) Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE). *Genome research,* **17** (6), 910–6.

[Billon & Côté, 2011] Billon, P. & Côté, J. (2011) Precise deposition of histone H2A.Z in chromatin for genome expression and maintenance. *Biochimica et biophysica acta*

[Blat & Kleckner, 1999] Blat, Y. & Kleckner, N. (1999) Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell,* **98** (2), 249–59.

[Boone *et al.*, 2010] Boone, N., Loriod, B., Bergon, A., Sbai, O., Formisano-Tréziny, C., Gabert, J., Khrestchatisky, M., Nguyen, C., Féron, F., Axelrod, F. B. & Ibrahim, E. C. (2010) Olfactory stem cells, a new cellular model for studying molecular mechanisms underlying familial dysautonomia. *PloS one,* **5** (12), e15590.

[Borgström *et al.*, 2011] Borgström, E., Lundin, S. & Lundeberg, J. (2011) Large scale library generation for high throughput sequencing. *PloS one,* **6** (4), e19119.

[Brazma *et al.*, 2001] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. & Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature genetics,* **29** (4), 365–71.

[Brazma *et al.*, 2003] Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., Rocca-Serra, P. & Sansone, S.-A. (2003) ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic acids research,* **31** (1), 68–71.

[Cai *et al.*, 2010] Cai, X., Hou, L., Su, N., Hu, H., Deng, M. & Li, X. (2010) Systematic identification of conserved motif modules in the human genome. *BMC genomics,* **11**, 567.

[Callow *et al.*, 2000] Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P. & Rubin, E. M. (2000) Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome research,* **10** (12), 2022–9.

[Carninci *et al.*, 2005] Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., de Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P. T., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. A. M., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J. & Hayashizaki, Y. (2005) The transcriptional landscape of the mammalian genome. *Science (New York, N.Y.),* **309** (5740), 1559–63.

[Chain *et al.*, 2010] Chain, B., Bowen, H., Hammond, J., Posch, W., Rasaiyaah, J., Tsang, J. & Noursadeghi, M. (2010) Error, reproducibility and sensitivity : a pipeline for data processing of Agilent oligonucleotide expression arrays. *BMC bioinformatics,* **11**, 344.

[Chan, 2005] Chan, E. Y. (2005) Advances in sequencing technology. *Mutation research,* **573** (1-2), 13–40.

[Chang *et al.*, 2011] Chang, H., Jackson, D. G., Kayne, P. S., Ross-Macdonald, P. B., Ryseck, R.-P. & Siemers, N. O. (2011) Exome Sequencing Reveals Comprehensive Genomic Alterations across Eight Cancer Cell Lines. *PloS one,* **6** (6), e21097.

[Chaouiya *et al.*, 2012] Chaouiya, C., Naldi, A. & Thieffry, D. (2012) Logical Modelling of Gene Regulatory Networks with GINsim. *Methods in molecular biology (Clifton, N.J.),* **804**, 463–79.

[Chen & Sharp, 2004] Chen, H. & Sharp, B. M. (2004) Content-rich biological network constructed by mining PubMed abstracts. *BMC bioinformatics,* **5**, 147.

[Chen & Sadowski, 2005] Chen, J. & Sadowski, I. (2005) Identification of the mismatch repair genes PMS2 and MLH1 as p53 target genes by using serial analysis of binding elements. *Proceedings of the National Academy of Sciences of the United States of America,* **102** (13), 4813–8.

[Chen *et al.*, 2008] Chen, R., Mallelwar, R., Thosar, A., Venkatasubrahmanyam, S. & Butte, A. J. (2008) GeneChaser : identifying all biological and clinical conditions in which genes of interest are differentially expressed. *BMC bioinformatics,* **9**, 548.

[Chiaretti *et al.*, 2004] Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J. & Foa, R. (2004) Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood,* **103** (7), 2771–8.

[Clark *et al.*, 2011] Clark, M. J., Chen, R., Lam, H. Y. K., Karczewski, K. J., Chen, R., Euskirchen, G., Butte, A. J. & Snyder, M. (2011) Performance comparison of exome DNA sequencing technologies. *Nature biotechnology*

[Cleveland, 1979] Cleveland, W. S. (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association,* **74** (368), 829–836.

[Core *et al.*, 2008] Core, L. J., Waterfall, J. J. & Lis, J. T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (New York, N.Y.),* **322** (5909), 1845–8.

[David *et al.*, 2011] David, M., Dzamba, M., Lister, D., Ilie, L. & Brudno, M. (2011) SHRiMP2 : sensitive yet practical SHort Read Mapping. *Bioinformatics (Oxford, England),* **27** (7), 1011–2.

[De Santa *et al.*, 2010] De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., Muller, H., Ragoussis, J., Wei, C.-L. & Natoli, G. (2010) A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS biology,* **8** (5), e1000384.

[Dekker *et al.*, 2002] Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. (2002) Capturing chromosome conformation. *Science (New York, N.Y.),* **295** (5558), 1306–11.

[Dobbin *et al.*, 2003] Dobbin, K., Shih, J. H. & Simon, R. (2003) Statistical design of reverse dye microarrays. *Bioinformatics (Oxford, England),* **19** (7), 803–10.

[Dohm *et al.*, 2008] Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research,* **36** (16), e105.

[Draghici *et al.*, 2003] Draghici, S., Kulaeva, O., Hoff, B., Petrov, A., Shams, S. & Tainsky, M. A. (2003) Noise sampling method : an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. *Bioinformatics (Oxford, England),* **19** (11), 1348–59.

[Droege & Hill, 2008] Droege, M. & Hill, B. (2008) The Genome Sequencer FLX System–longer reads, more applications, straight forward bioinformatics and more complete data sets. *Journal of biotechnology,* **136** (1-2), 3–10.

[Dunn *et al.*, 2007] Dunn, J. J., McCorkle, S. R., Everett, L. & Anderson, C. W. (2007) Paired-end genomic signature tags : a method for the functional analysis of genomes and epigenomes. *Genetic engineering,* **28**, 159–73.

[Edgar *et al.*, 2002] Edgar, R., Domrachev, M. & Lash, A. E. (2002) Gene Expression Omnibus : NCBI gene expression and hybridization array data repository. *Nucleic acids research,* **30** (1), 207–10.

[Eisen *et al.*, 1998] Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America,* **95** (25), 14863–8.

[Elnitski *et al.*, 2006] Elnitski, L., Jin, V. X., Farnham, P. J. & Jones, S. J. M. (2006) Locating mammalian transcription factor binding sites : a survey of computational and experimental techniques. *Genome research,* **16** (12), 1455–64.

[Enright *et al.*, 2002] Enright, A. J., Van Dongen, S. & Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic acids research,* **30** (7), 1575–84.

[Ewing & Green, 1998] Ewing, B. & Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research,* **8** (3), 186–94.

[Ewing *et al.*, 1998] Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research,* **8** (3), 175–85.

[Fedorova & Zink, 2008] Fedorova, E. & Zink, D. (2008) Nuclear architecture and gene regulation. *Biochimica et biophysica acta,* **1783** (11), 2174–84.

[Feng *et al.*, 2009] Feng, C., Araki, M., Kunimoto, R., Tamon, A., Makiguchi, H., Niijima, S., Tsujimoto, G. & Okuno, Y. (2009) GEM-TREND : a web tool for gene expression data mining toward relevant network discovery. *BMC genomics,* **10**, 411.

[Ferdin *et al.*, 2010] Ferdin, J., Kunej, T. & Calin, G. A. (2010) Non-coding RNAs : identification of cancer-associated microRNAs by gene profiling. *Technology in cancer research & treatment,* **9** (2), 123–38.

[Fernandez-Capetillo *et al.*, 2003] Fernandez-Capetillo, O., Mahadevaiah, S. K., Celeste, A., Romanienko, P. J., Camerini-Otero, R. D., Bonner, W. M., Manova, K., Burgoyne, P. & Nussenzweig, A. (2003) H2AX is required for chromatin remodeling and inactivation of sex chromosomes in male mouse meiosis. *Developmental cell,* **4** (4), 497–508.

[Fiume *et al.*, 2010] Fiume, M., Williams, V., Brook, A. & Brudno, M. (2010) Savant : genome browser for high-throughput sequencing data. *Bioinformatics (Oxford, England),* **26** (16), 1938–44.

[Foltz *et al.*, 2009] Foltz, D. R., Jansen, L. E. T., Bailey, A. O., Yates, J. R., Bassett, E. A., Wood, S., Black, B. E. & Cleveland, D. W. (2009) Centromere-specific assembly of CENP-a nucleosomes is mediated by HJURP. *Cell,* **137** (3), 472–84.

[Fullwood *et al.*, 2009] Fullwood, M. J., Wei, C.-L., Liu, E. T. & Ruan, Y. (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome research,* **19** (4), 521–32.

[Geyer & Corces, 1992] Geyer, P. K. & Corces, V. G. (1992) DNA position-specific repression of transcription by a Drosophila zinc finger protein. *Genes & development,* **6** (10), 1865–73.

[Gheldof *et al.*, 2012] Gheldof, N., Leleu, M., Noordermeer, D., Rougemont, J. & Reymond, A. (2012) Detecting long-range chromatin interactions using the chromosome conformation capture sequencing (4C-seq) method. *Methods in molecular biology (Clifton, N.J.),* **786**, 211–25.

[Giardine *et al.*, 2005] Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J. & Nekrutenko, A. (2005) Galaxy : a platform for interactive large-scale genome analysis. *Genome research,* **15** (10), 1451–5.

[Gilmour & Lis, 1985] Gilmour, D. S. & Lis, J. T. (1985) In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster. *Molecular and cellular biology,* **5** (8), 2009–18.

[Giresi *et al.*, 2007] Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research,* **17** (6), 877–85.

[Glenn, 2011] Glenn, T. C. (2011) Field guide to next-generation DNA sequencers. *Molecular ecology resources*

[Goldfeder *et al.*, 2011] Goldfeder, R. L., Parker, S. C. J., Ajay, S. S., Ozel Abaan, H. & Margulies, E. H. (2011) A bioinformatics approach for determining sample identity from different lanes of high-throughput sequencing data. *PloS one,* **6** (8), e23683.

[Golub *et al.*, 1999] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S. (1999) Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science (New York, N.Y.),* **286** (5439), 531–7.

[Gommans & Berezikov, 2012] Gommans, W. M. & Berezikov, E. (2012) Sample preparation for small RNA massive parallel sequencing. *Methods in molecular biology (Clifton, N.J.),* **786**, 167–78.

[Good *et al.*, 2006] Good, B. M., Kawas, E. A., Kuo, B. Y.-L. & Wilkinson, M. D. (2006) iHOPerator : user-scripting a personalized bioinformatics Web, starting with the iHOP website. *BMC bioinformatics,* **7**, 534.

[Govin *et al.*, 2004] Govin, J., Caron, C., Lestrat, C., Rousseaux, S. & Khochbin, S. (2004) The role of histones in chromatin remodelling during mammalian spermiogenesis. *European journal of biochemistry / FEBS,* **271** (17), 3459–69.

[Guttman *et al.*, 2009] Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen,

T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L. & Lander, E. S. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature,* **458** (7235), 223–7.

[Hansen *et al.*, 2011] Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R. A. & Feinberg, A. P. (2011) Increased methylation variation in epigenetic domains across cancer types. *Nature genetics*

[Hatzis *et al.*, 2008] Hatzis, P., van der Flier, L. G., van Driel, M. A., Guryev, V., Nielsen, F., Denissov, S., Nijman, I. J., Koster, J., Santo, E. E., Welboren, W., Versteeg, R., Cuppen, E., van de Wetering, M., Clevers, H. & Stunnenberg, H. G. (2008) Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. *Molecular and cellular biology,* **28** (8), 2732–44.

[Heintzman *et al.*, 2009] Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M. & Ren, B. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature,* **459** (7243), 108–12.

[Hillier *et al.*, 2008] Hillier, L. W., Marth, G. T., Quinlan, A. R., Dooling, D., Fewell, G., Barnett, D., Fox, P., Glasscock, J. I., Hickenbotham, M., Huang, W., Magrini, V. J., Richt, R. J., Sander, S. N., Stewart, D. A., Stromberg, M., Tsung, E. F., Wylie, T., Schedl, T., Wilson, R. K. & Mardis, E. R. (2008) Whole-genome sequencing and variant discovery in C. elegans. *Nature methods,* **5** (2), 183–8.

[Hillmer *et al.*, 2011] Hillmer, A. M., Yao, F., Inaki, K., Lee, W. H., Ariyaratne, P. N., Teo, A. S. M., Woo, X. Y., Zhang, Z., Zhao, H., Ukil, L., Chen, J. P., Zhu, F., So, J. B. Y., Salto-Tellez, M., Poh, W. T., Zawack, K. F. B., Nagarajan, N., Gao, S., Li, G., Kumar, V., Lim, H. P. J., Sia, Y. Y., Chan, C. S., Leong, S. T., Neo, S. C., Choi, P. S. D., Thoreau, H., Tan, P. B. O., Shahab, A., Ruan, X., Bergh, J., Hall, P., Cacheux-Rataboul, V., Wei, C.-L., Yeoh, K. G., Sung, W.-K., Bourque, G., Liu, E. T. & Ruan, Y. (2011) Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. *Genome research,* **21** (5), 665–75.

[Hims *et al.*, 2007] Hims, M. M., Shetty, R. S., Pickel, J., Mull, J., Leyne, M., Liu, L., Gusella, J. F. & Slaugenhaupt, S. A. (2007) A humanized IKBKAP transgenic mouse models a tissue-specific human splicing defect. *Genomics,* **90** (3), 389–96.

[Ho *et al.*, 2011] Ho, J. W. K., Bishop, E., Karchenko, P. V., Nègre, N., White, K. P. & Park, P. J. (2011) ChIP-chip versus ChIP-seq : lessons for experimental design and data analysis. *BMC genomics,* **12**, 134.

[Holt & Jones, 2008] Holt, R. A. & Jones, S. J. M. (2008) The new paradigm of flow cell sequencing. *Genome research,* **18** (6), 839–46.

[Holtgrewe *et al.*, 2011] Holtgrewe, M., Emde, A.-K., Weese, D. & Reinert, K. (2011) A novel and well-defined benchmarking method for second generation read mapping. *BMC bioinformatics,* **12**, 210.

[Homer *et al.*, 2009] Homer, N., Merriman, B. & Nelson, S. F. (2009) BFAST : an alignment tool for large scale genome resequencing. *PloS one,* **4** (11), e7767.

[Hong *et al.*, 2011] Hong, L. Z., Li, J., Schmidt-Küntzel, A., Warren, W. C. & Barsh, G. S. (2011) Digital gene expression for non-model organisms. *Genome research,* **21** (11), 1905–15.

[Hruz *et al.*, 2008] Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W. & Zimmermann, P. (2008) Genevestigator v3 : a reference expression database for the meta-analysis of transcriptomes. *Advances in bioinformatics,* **2008**, 420747.

[Huang *et al.*, 2009] Huang, D. W., Sherman, B. T. & Lempicki, R. A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols,* **4** (1), 44–57.

[Hurtado *et al.*, 2011] Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. & Carroll, J. S. (2011) FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nature genetics,* **43** (1), 27–33.

[Hutchison, 2007] Hutchison, C. A. (2007) DNA sequencing : bench to bedside and beyond. *Nucleic acids research,* **35** (18), 6227–37.

[Hyman, 1988] Hyman, E. D. (1988) A new method of sequencing DNA. *Analytical biochemistry,* **174** (2), 423–36.

[Impey *et al.*, 2004] Impey, S., McCorkle, S. R., Cha-Molstad, H., Dwyer, J. M., Yochum, G. S., Boss, J. M., McWeeney, S., Dunn, J. J., Mandel, G. & Goodman, R. H. (2004) Defining the CREB regulon : a genome-wide analysis of transcription factor regulatory regions. *Cell,* **119** (7), 1041–54.

[Inza *et al.*, 2004] Inza, I. n., Larrañaga, P., Blanco, R. & Cerrolaza, A. J. (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial intelligence in medicine,* **31** (2), 91–103.

[Irizarry *et al.*, 2005] Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G. N., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S. C., Hoffman, E., Jedlicka, A. E., Kawasaki, E., Martínez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S. Q. & Yu, W. (2005) Multiple-laboratory comparison of microarray platforms. *Nature methods,* **2** (5), 345–50.

[Jenuwein & Allis, 2001] Jenuwein, T. & Allis, C. D. (2001) Translating the histone code. *Science (New York, N.Y.),* **293** (5532), 1074–80.

[Johnson *et al.*, 2007] Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.),* **316** (5830), 1497–502.

[Kaikkonen *et al.*, 2011] Kaikkonen, M. U., Lam, M. T. Y. & Glass, C. K. (2011) Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovascular research,* **90** (3), 430–40.

[Kauffmann & Huber, 2008] Kauffmann, A. & Huber, W. (2008) *arrayQualityMetrics : Quality metrics on microarray data sets.* R package version 2.4.3.

[Kauffmann *et al.*, 2009] Kauffmann, A., Rayner, T. F., Parkinson, H., Kapushesky, M., Lukk, M., Brazma, A. & Huber, W. (2009) Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics (Oxford, England),* **25** (16), 2092–4.

[Kellum & Schedl, 1992] Kellum, R. & Schedl, P. (1992) A group of scs elements function as domain boundaries in an enhancer-blocking assay. *Molecular and cellular biology,* **12** (5), 2424–31.

[Keren *et al.*, 2010] Keren, H., Donyo, M., Zeevi, D., Maayan, C., Pupko, T. & Ast, G. (2010) Phosphatidylserine increases IKBKAP levels in familial dysautonomia cells. *PloS one,* **5** (12), e15884.

[Khalil *et al.*, 2009] Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B. E., van Oudenaarden, A., Regev, A., Lander, E. S. & Rinn, J. L. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America,* **106** (28), 11667–72.

[Kharchenko *et al.*, 2008] Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology,* **26** (12), 1351–9.

[Kidder *et al.*, 2011] Kidder, B. L., Hu, G. & Zhao, K. (2011) ChIP-Seq : technical considerations for obtaining high-quality data. *Nature immunology,* **12** (10), 918–22.

[Kim *et al.*, 2008] Kim, C., Cheon, M., Kang, M. & Chang, I. (2008) A simple and exact Laplacian clustering of complex networking phenomena : application to gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America,* **105** (11), 4083–7.

[Kim *et al.*, 2010] Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G. & Greenberg, M. E. (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature,* **465** (7295), 182–7.

[Kircher *et al.*, 2011] Kircher, M., Heyn, P. & Kelso, J. (2011) Addressing challenges in the production and analysis of illumina sequencing data. *BMC genomics,* **12**, 382.

[Kircher *et al.*, 2009] Kircher, M., Stenzel, U. & Kelso, J. (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome biology,* **10** (8), R83.

[Klose & Zhang, 2007] Klose, R. J. & Zhang, Y. (2007) Regulation of histone methylation by demethylimination and demethylation. *Nature reviews. Molecular cell biology,* **8** (4), 307–18.

[Knapen *et al.*, 2009] Knapen, D., Vergauwen, L., Laukens, K. & Blust, R. (2009) Best practices for hybridization design in two-colour microarray analysis. *Trends in biotechnology,* **27** (7), 406–14.

[Kohonen, 1997] Kohonen, T. (1997) Self-Organizing Maps. *New York : Springer*

[Korbel *et al.*, 2007] Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C. E., Chi, J., Yang, F., Carter, N. P., Hurles, M. E., Weissman, S. M., Harkins, T. T., Gerstein, M. B., Egholm, M. & Snyder, M. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, N.Y.),* **318** (5849), 420–6.

[Kornberg, 1999] Kornberg, R. D. (1999) Eukaryotic transcriptional control. *Trends in cell biology,* **9** (12), M46–9.

[Kouzarides, 2007] Kouzarides, T. (2007) Chromatin modifications and their function. *Cell,* **128** (4), 693–705.

[Krol *et al.*, 2010] Krol, J., Loedige, I. & Filipowicz, W. (2010) The widespread regulation of microRNA biogenesis, function and decay. *Nature reviews. Genetics,* **11** (9), 597–610.

[Lacoste & Côté, 2003] Lacoste, N. & Côté, J. (2003) [The epigenetic code of histones]. *Médecine sciences : M/S,* **19** (10), 955–9.

[Lan *et al.*, 2011] Lan, X., Adams, C., Landers, M., Dudas, M., Krissinger, D., Marnellos, G., Bonneville, R., Xu, M., Wang, J., Huang, T. H.-M., Meredith, G. & Jin, V. X. (2011) High resolution detection and analysis of CpG dinucleotides methylation using MBD-Seq technology. *PloS one,* **6** (7), e22226.

[Langmead *et al.*, 2009] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology,* **10** (3), R25.

[Lee & Mahadevan, 2009] Lee, B. M. & Mahadevan, L. C. (2009) Stability of histone modifications across mammalian genomes : implications for 'epigenetic' marking. *Journal of cellular biochemistry,* **108** (1), 22–34.

[Lee & Young, 2000] Lee, T. I. & Young, R. A. (2000) Transcription of eukaryotic protein-coding genes. *Annual review of genetics,* **34**, 77–137.

[Lemoine *et al.*, 2006] Lemoine, S., Combes, F., Servant, N. & Le Crom, S. (2006) Goulphar : rapid access and expertise for standard two-color microarray normalization methods. *BMC bioinformatics,* **7**, 467.

[Li *et al.*, 2007] Li, B., Carey, M. & Workman, J. L. (2007) The role of chromatin during transcription. *Cell,* **128** (4), 707–19.

[Li & Durbin, 2009] Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England),* **25** (14), 1754–60.

[Li *et al.*, 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England),* **25** (16), 2078–9.

[Li & Homer, 2010] Li, H. & Homer, N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics,* **11** (5), 473–83.

[Li *et al.*, 2011] Li, J., Zhao, Q. & Bolund, L. (2011) Computational methods for epigenetic analysis : the protocol of computational analysis for modified methylation-specific digital karyotyping based on massively parallel sequencing. *Methods in molecular biology (Clifton, N.J.),* **791**, 313–28.

[Linsen *et al.*, 2009] Linsen, S. E. V., de Wit, E., Janssens, G., Heater, S., Chapman, L., Parkin, R. K., Fritz, B., Wyman, S. K., de Bruijn, E., Voest, E. E., Kuersten, S., Tewari, M. & Cuppen, E. (2009) Limitations and possibilities of small RNA digital gene expression profiling. *Nature methods,* **6** (7), 474–6.

[Loots & Ovcharenko, 2007] Loots, G. & Ovcharenko, I. (2007) ECRbase : database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. *Bioinformatics (Oxford, England),* **23** (1), 122–4.

[Lopez *et al.*, 2008] Lopez, F., Textoris, J., Bergon, A., Didier, G., Remy, E., Granjeaud, S., Imbert, J., Nguyen, C. & Puthier, D. (2008) TranscriptomeBrowser : a powerful and flexible toolbox to explore productively the transcriptional landscape of the Gene Expression Omnibus database. *PloS one,* **3** (12), e4001.

[Lopez-Romero, 2008] Lopez-Romero, P. (2008) *Agi4x44PreProcess : PreProcessing of Agilent 4x44 array data.* R package version 1.6.0.

[Lu *et al.*, 2010] Lu, F., Wikramasinghe, P., Norseen, J., Tsai, K., Wang, P., Showe, L., Davuluri, R. V. & Lieberman, P. M. (2010) Genome-wide analysis of host-chromosome binding sites for Epstein-Barr Virus Nuclear Antigen 1 (EBNA1). *Virology journal,* **7**, 262.

[Machanick & Bailey, 2011] Machanick, P. & Bailey, T. L. (2011) MEME-ChIP : motif analysis of large DNA datasets. *Bioinformatics (Oxford, England),* **27** (12), 1696–7.

[Maher *et al.*, 2009] Maher, C. A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N. & Chinnaiyan, A. M. (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature,* **458** (7234), 97–101.

[Mardis, 2007] Mardis, E. R. (2007) ChIP-seq : welcome to the new frontier. *Nature methods,* **4** (8), 613–4.

[Margulies *et al.*, 2005] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. & Rothberg, J. M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* **437** (7057), 376–80.

[Martens-Uzunova *et al.*, 2011] Martens-Uzunova, E. S., Jalava, S. E., Dits, N. F., van Leenders, G. J. L. H., Mø ller, S., Trapman, J., Bangma, C. H., Litman, T., Visakorpi, T. & Jenster, G. (2011) Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. *Oncogene*

[Martin *et al.*, 2004] Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D. & Jacq, B. (2004) GOToolBox : functional analysis of gene datasets based on Gene Ontology. *Genome biology,* **5** (12), R101.

[Massie & Mills, 2008] Massie, C. E. & Mills, I. G. (2008) ChIPping away at gene regulation. *EMBO reports,* **9** (4), 337–43.

[Matys *et al.*, 2003] Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D.-U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. & Wingender, E. (2003) TRANSFAC : transcriptional regulation, from patterns to profiles. *Nucleic acids research,* **31** (1), 374–8.

[Maxam & Gilbert, 1977] Maxam, A. M. & Gilbert, W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America,* **74** (2), 560–4.

[McKenna *et al.*, 2010] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. A. (2010) The Genome Analysis Toolkit : a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research,* **20** (9), 1297–303.

[McLean *et al.*, 2010] McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M. & Bejerano, G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology,* **28** (5), 495–501.

[Metzker, 2010] Metzker, M. L. (2010) Sequencing technologies - the next generation. *Nature reviews. Genetics,* **11** (1), 31–46.

[Meyerson *et al.*, 2010] Meyerson, M., Gabriel, S. & Getz, G. (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nature reviews. Genetics,* **11** (10), 685–96.

[Moorthy & Mohamad, 2011] Moorthy, K. & Mohamad, M. S. (2011) Random forest for gene selection and microarray data classification. *Bioinformation,* **7** (3), 142–6.

[Morin *et al.*, 2008] Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S. & Marra, M. (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques,* **45** (1), 81–94.

[Morris *et al.*, 2008] Morris, K. V., Santoso, S., Turner, A.-M., Pastori, C. & Hawkins, P. G. (2008) Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. *PLoS genetics,* **4** (11), e1000258.

[Mutter & Boynton, 1995] Mutter, G. L. & Boynton, K. A. (1995) PCR bias in amplification of androgen receptor alleles, a trinucleotide repeat marker used in clonality studies. *Nucleic acids research,* **23** (8), 1411–8.

[Mutter *et al.*, 2004] Mutter, G. L., Zahrieh, D., Liu, C., Neuberg, D., Finkelstein, D., Baker, H. E. & Warrington, J. A. (2004) Comparison of frozen and RNALater solid tissue storage methods for use in RNA expression microarrays. *BMC genomics,* **5**, 88.

[Naef & Huelsken, 2005] Naef, F. & Huelsken, J. (2005) Cell-type-specific transcriptomics in chimeric models using transcriptome-based masks. *Nucleic acids research,* **33** (13), e111.

[Nammo *et al.*, 2011] Nammo, T., Rodríguez-Seguí, S. A. & Ferrer, J. (2011) Mapping open chromatin with formaldehyde-assisted isolation of regulatory elements. *Methods in molecular biology (Clifton, N.J.),* **791**, 287–96.

[Narurkar *et al.*, 1968] Narurkar, M. V., Narurkar, L. M. & Sahasrabudhe, M. B. (1968) A new technique of pH gradient electrophoresis as applied to the separation of nucleic acid bases. *Analytical biochemistry,* **26** (1), 174–7.

[Natoli, 2011] Natoli, G. (2011) Specialized chromatin patterns in the control of inflammatory gene expression. *Current topics in microbiology and immunology,* **349**, 61–72.

[Ng *et al.*, 2005] Ng, P., Wei, C.-L., Sung, W.-K., Chiu, K. P., Lipovich, L., Ang, C. C., Gupta, S., Shahab, A., Ridwan, A., Wong, C. H., Liu, E. T. & Ruan, Y. (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature methods,* **2** (2), 105–11.

[Ø rom *et al.*, 2010] Ø rom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., Guigo, R. & Shiekhattar, R. (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell,* **143** (1), 46–58.

[Obayashi & Kinoshita, 2011] Obayashi, T. & Kinoshita, K. (2011) COXPRESdb : a database to compare gene coexpression in seven model animals. *Nucleic acids research,* **39** (Database issue), D1016–22.

[Oberthuer *et al.*, 2010] Oberthuer, A., Juraeva, D., Li, L., Kahlert, Y., Westermann, F., Eils, R., Berthold, F., Shi, L., Wolfinger, R. D., Fischer, M. & Brors, B. (2010) Comparison of performance of one-color and two-color gene-expression analyses in predicting clinical endpoints of neuroblastoma patients. *The pharmacogenomics journal,* **10** (4), 258–66.

[O'Geen *et al.*, 2011] O'Geen, H., Echipare, L. & Farnham, P. J. (2011) Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications. *Methods in molecular biology (Clifton, N.J.),* **791**, 265–86.

[Okada *et al.*, 2005] Okada, T., Endo, M., Singh, M. B. & Bhalla, P. L. (2005) Analysis of the histone H3 gene family in Arabidopsis and identification of the male-gamete-specific variant AtMGH3. *The Plant journal : for cell and molecular biology,* **44** (4), 557–68.

[Olguin-Lamas *et al.*, 2011] Olguin-Lamas, A., Madec, E., Hovasse, A., Werkmeister, E., Callebaut, I., Slomianny, C., Delhaye, S., Mouveaux, T., Schaeffer-Reiss, C., Van Dorsselaer, A. & Tomavo, S. (2011) A novel Toxoplasma gondii nuclear factor TgNF3 is a dynamic chromatin-associated component, modulator of nucleolar architecture and parasite virulence. *PLoS pathogens,* **7** (3), e1001328.

[Orphanides *et al.*, 1996] Orphanides, G., Lagrange, T. & Reinberg, D. (1996) The general transcription factors of RNA polymerase II. *Genes & development,* **10** (21), 2657–83.

[Paquet & Yang, 2008] Paquet, A. & Yang, J. Y. H. (2008) *arrayQuality : Assessing array quality on spotted arrays*. R package version 1.24.0.

[Pareek *et al.*, 2011] Pareek, C. S., Smoczynski, R. & Tretyn, A. (2011) Sequencing technologies and genome sequencing. *Journal of applied genetics*

[Parkinson *et al.*, 2011] Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Holloway, E., Kurbatova, N., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rustici, G., Sharma, A., Williams, E., Adamusiak, T., Brandizi, M., Sklyar, N. & Brazma, A. (2011) ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic acids research,* **39** (Database issue), D1002–4.

[Pastinen *et al.*, 2000] Pastinen, T., Raitio, M., Lindroos, K., Tainola, P., Peltonen, L. & Syvänen, A. C. (2000) A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. *Genome research,* **10** (7), 1031–42.

[Pekowska *et al.*, 2010] Pekowska, A., Benoukraf, T., Ferrier, P. & Spicuglia, S. (2010) A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome research,* **20** (11), 1493–502.

[Pepke *et al.*, 2009] Pepke, S., Wold, B. & Mortazavi, A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nature methods,* **6** (11 Suppl), S22–32.

[Perou *et al.*, 2000] Perou, C. M., Sø rlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lø nning, P. E., Bø rresen Dale, A. L., Brown, P. O. & Botstein, D. (2000) Molecular portraits of human breast tumours. *Nature,* **406** (6797), 747–52.

[Place *et al.*, 2008] Place, R. F., Li, L.-C., Pookot, D., Noonan, E. J. & Dahiya, R. (2008) MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proceedings of the National Academy of Sciences of the United States of America,* **105** (5), 1608–13.

[Ponting *et al.*, 2009] Ponting, C. P., Oliver, P. L. & Reik, W. (2009) Evolution and functions of long noncoding RNAs. *Cell,* **136** (4), 629–41.

[Prober *et al.*, 1987] Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., Jensen, M. A. & Baumeister, K. (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science (New York, N.Y.),* **238** (4825), 336–41.

[Quinlan & Hall, 2010] Quinlan, A. R. & Hall, I. M. (2010) BEDTools : a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England),* **26** (6), 841–2.

[Ransohoff & Gourlay, 2010] Ransohoff, D. F. & Gourlay, M. L. (2010) Sources of bias in specimens for research about molecular markers for cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* **28** (4), 698–704.

[Ravasi *et al.*, 2010] Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., Carninci, P., Daub, C. O., Forrest, A. R. R., Gough, J., Grimmond, S., Han, J.-H., Hashimoto, T., Hide, W., Hofmann, O., Kamburov, A., Kaur, M., Kawaji, H., Kubosaki, A., Lassmann, T., van Nimwegen, E., MacPherson, C. R., Ogawa, C., Radovanovic, A., Schwartz, A., Teasdale, R. D., Tegnér, J., Lenhard, B., Teichmann, S. A., Arakawa, T., Ninomiya, N., Murakami, K., Tagami, M., Fukuda, S., Imamura, K., Kai, C., Ishihara, R., Kitazume, Y., Kawai, J., Hume, D. A., Ideker, T. & Hayashizaki, Y. (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell,* **140** (5), 744–52.

[Ren *et al.*, 2000] Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P. & Young, R. A. (2000) Genome-wide location and function of DNA binding proteins. *Science (New York, N.Y.),* **290** (5500), 2306–9.

[Robyr *et al.*, 2002] Robyr, D., Suka, Y., Xenarios, I., Kurdistani, S. K., Wang, A., Suka, N. & Grunstein, M. (2002) Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell,* **109** (4), 437–46.

[Roh *et al.*, 2004] Roh, T.-y., Ngau, W. C., Cui, K., Landsman, D. & Zhao, K. (2004) High-resolution genome-wide mapping of histone modifications. *Nature biotechnology,* **22** (8), 1013–6.

[Ronaghi, 2001] Ronaghi, M. (2001) Pyrosequencing sheds light on DNA sequencing. *Genome research,* **11** (1), 3–11.

[Ronaghi *et al.*, 1998] Ronaghi, M., Uhlén, M. & Nyrén, P. (1998) A sequencing method based on real-time pyrophosphate. *Science (New York, N.Y.),* **281** (5375), 363, 365.

[Ross & Cronin, 2011] Ross, J. S. & Cronin, M. (2011) Whole cancer genome sequencing by next-generation methods. *American journal of clinical pathology,* **136** (4), 527–39.

[Ross-Innes *et al.*, 2011] Ross-Innes, C. S., Brown, G. D. & Carroll, J. S. (2011) A co-ordinated interaction between CTCF and ER in breast cancer cells. *BMC genomics,* **12** (1), 593.

[Rothberg *et al.*, 2011] Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., McKernan, K. J., Williams, A., Roth, G. T. & Bustillo, J. (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature,* **475** (7356), 348–52.

[Rothberg & Leamon, 2008] Rothberg, J. M. & Leamon, J. H. (2008) The development and impact of 454 sequencing. *Nature biotechnology,* **26** (10), 1117–24.

[Rougemont *et al.*, 2008] Rougemont, J., Amzallag, A., Iseli, C., Farinelli, L., Xenarios, I. & Naef, F. (2008) Probabilistic base calling of Solexa sequencing data. *BMC bioinformatics,* **9**, 431.

[Rumble *et al.*, 2009] Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A. & Brudno, M. (2009) SHRiMP : accurate mapping of short color-space reads. *PLoS computational biology,* **5** (5), e1000386.

[Rye *et al.*, 2011] Rye, M. B., Sæ trom, P. l. & Drablø s, F. (2011) A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic acids research,* **39** (4), e25.

[Sanger *et al.*, 1977] Sanger, F., Nicklen, S. & Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America,* **74** (12), 5463–7.

[Schones & Zhao, 2008] Schones, D. E. & Zhao, K. (2008) Genome-wide approaches to studying chromatin modifications. *Nature reviews. Genetics,* **9** (3), 179–91.

[Sean & Meltzer, 2007] Sean, D. & Meltzer, P. S. (2007) GEOquery : a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics (Oxford, England),* **23** (14), 1846–7.

[Shendure *et al.*, 2005] Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. & Church, G. M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.),* **309** (5741), 1728–32.

[Shi *et al.*, 2010] Shi, L., Campbell, G., Jones, W. D., Campagne, F., Wen, Z., Walker, S. J., Su, Z., Chu, T.-M., Goodsaid, F. M., Pusztai, L., Shaughnessy, J. D., Oberthuer, A., Thomas, R. S., Paules, R. S., Fielden, M., Barlogie, B., Chen, W., Du, P., Fischer, M., Furlanello, C., Gallas, B. D., Ge, X., Megherbi, D. B., Symmans, W. F., Wang, M. D., Zhang, J., Bitter, H., Brors, B., Bushel, P. R., Bylesjo, M., Chen, M., Cheng, J., Cheng, J., Chou, J., Davison, T. S.,

Delorenzi, M., Deng, Y., Devanarayan, V., Dix, D. J., Dopazo, J., Dorff, K. C., Elloumi, F., Fan, J., Fan, S., Fan, X., Fang, H., Gonzaludo, N., Hess, K. R., Hong, H., Huan, J., Irizarry, R. A., Judson, R., Juraeva, D., Lababidi, S., Lambert, C. G., Li, L., Li, Y., Li, Z., Lin, S. M., Liu, G., Lobenhofer, E. K., Luo, J., Luo, W., McCall, M. N., Nikolsky, Y., Pennello, G. A., Perkins, R. G., Philip, R., Popovici, V., Price, N. D., Qian, F., Scherer, A., Shi, T., Shi, W., Sung, J., Thierry-Mieg, D., Thierry-Mieg, J., Thodima, V., Trygg, J., Vishnuvajjala, L., Wang, S. J., Wu, J., Wu, Y., Xie, Q., Yousef, W. A., Zhang, L., Zhang, X., Zhong, S., Zhou, Y., Zhu, S., Arasappan, D., Bao, W., Lucas, A. B., Berthold, F., Brennan, R. J., Buness, A., Catalano, J. G., Chang, C., Chen, R., Cheng, Y., Cui, J., Czika, W., Demichelis, F., Deng, X., Dosymbekov, D., Eils, R., Feng, Y., Fostel, J., Fulmer-Smentek, S., Fuscoe, J. C., Gatto, L., Ge, W., Goldstein, D. R., Guo, L., Halbert, D. N., Han, J., Harris, S. C., Hatzis, C., Herman, D., Huang, J., Jensen, R. V., Jiang, R., Johnson, C. D., Jurman, G., Kahlert, Y., Khuder, S. A., Kohl, M., Li, J., Li, L., Li, M., Li, Q.-Z., Li, S., Li, Z., Liu, J., Liu, Y., Liu, Z., Meng, L., Madera, M., Martinez-Murillo, F., Medina, I., Meehan, J., Miclaus, K., Moffitt, R. A., Montaner, D., Mukherjee, P., Mulligan, G. J., Neville, P., Nikolskaya, T., Ning, B., Page, G. P., Parker, J., Parry, R. M., Peng, X., Peterson, R. L., Phan, J. H., Quanz, B., Ren, Y., Riccadonna, S., Roter, A. H., Samuelson, F. W., Schumacher, M. M., Shambaugh, J. D., Shi, Q., Shippy, R., Si, S., Smalter, A., Sotiriou, C., Soukup, M., Staedtler, F., Steiner, G., Stokes, T. H., Sun, Q., Tan, P.-Y., Tang, R., Tezak, Z., Thorn, B., Tsyganova, M., Turpaz, Y., Vega, S. C., Visintainer, R., von Frese, J., Wang, C., Wang, E., Wang, J., Wang, W., Westermann, F., Willey, J. C., Woods, M., Wu, S., Xiao, N., Xu, J., Xu, L., Yang, L., Zeng, X., Zhang, J., Zhang, L., Zhang, M., Zhao, C., Puri, R. K., Scherf, U., Tong, W. & Wolfinger, R. D. (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology,* **28** (8), 827–38.

[Shi *et al.*, 2006] Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A., Willey, J. C., Setterquist, R. A., Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., Goodsaid, F. M., Herman, D., Jensen, R. V., Johnson, C. D., Lobenhofer, E. K., Puri, R. K., Schrf, U., Thierry-Mieg, J., Wang, C., Wilson, M., Wolber, P. K., Zhang, L., Amur, S., Bao, W., Barbacioru, C. C., Lucas, A. B., Bertholet, V., Boysen, C., Bromley, B., Brown, D., Brunner, A., Canales, R., Cao, X. M., Cebula, T. A., Chen, J. J., Cheng, J., Chu, T.-M., Chudin, E., Corson, J., Corton, J. C., Croner, L. J., Davies, C., Davison, T. S., Delenstarr, G., Deng, X., Dorris, D., Eklund, A. C., Fan, X.-h., Fang, H., Fulmer-Smentek, S., Fuscoe, J. C., Gallagher, K., Ge, W., Guo, L., Guo, X., Hager, J., Haje, P. K., Han, J., Han, T., Harbottle, H. C., Harris, S. C., Hatchwell, E., Hauser, C. A., Hester, S., Hong, H., Hurban, P., Jackson, S. A., Ji, H., Knight, C. R., Kuo, W. P., LeClerc, J. E., Levy, S., Li, Q.-Z., Liu, C., Liu, Y., Lombardi, M. J., Ma, Y., Magnuson, S. R., Maqsodi, B., McDaniel, T., Mei, N., Myklebost, O., Ning, B., Novoradovskaya, N., Orr, M. S., Osborn, T. W., Papallo, A., Patterson, T. A., Perkins, R. G., Peters, E. H., Peterson, R., Philips, K. L., Pine, P. S., Pusztai, L., Qian, F., Ren, H., Rosen, M., Rosenzweig, B. A., Samaha, R. R., Schena, M., Schroth, G. P., Shchegrova, S., Smith, D. D., Staedtler, F., Su, Z., Sun, H., Szallasi, Z., Tezak, Z., Thierry-Mieg, D., Thompson, K. L., Tikhonova, I., Turpaz, Y., Vallanat, B., Van, C., Walker, S. J., Wang, S. J., Wang, Y., Wolfinger, R., Wong, A., Wu, J., Xiao, C., Xie, Q., Xu, J., Yang, W., Zhang, L.,

Zhong, S., Zong, Y. & Slikker, W. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature biotechnology,* **24** (9), 1151–61.

[Shin *et al.*, 2009] Shin, H., Liu, T., Manrai, A. K. & Liu, X. S. (2009) CEAS : cis-regulatory element annotation system. *Bioinformatics (Oxford, England),* **25** (19), 2605–6.

[Shinde *et al.*, 2010] Shinde, K., Phatak, M., Johannes, F. M., Chen, J., Li, Q., Vineet, J. K., Hu, Z., Ghosh, K., Meller, J. & Medvedovic, M. (2010) Genomics Portals : integrative web-platform for mining genomics data. *BMC genomics,* **11**, 27.

[Siddiqui *et al.*, 2006] Siddiqui, A. S., Delaney, A. D., Schnerch, A., Griffith, O. L., Jones, S. J. M. & Marra, M. A. (2006) Sequence biases in large scale gene expression profiling data. *Nucleic acids research,* **34** (12), e83.

[Sims *et al.*, 2004] Sims, R. J., Mandal, S. S. & Reinberg, D. (2004) Recent highlights of RNA-polymerase-II-mediated transcription. *Current opinion in cell biology,* **16** (3), 263–71.

[Slaugenhaupt *et al.*, 2004] Slaugenhaupt, S. A., Mull, J., Leyne, M., Cuajungco, M. P., Gill, S. P., Hims, M. M., Quintero, F., Axelrod, F. B. & Gusella, J. F. (2004) Rescue of a human mRNA splicing defect by the plant cytokinin kinetin. *Human molecular genetics,* **13** (4), 429–36.

[Smyth, 2005] Smyth, G. K. (2005) Limma : linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, (Gentleman, R., Carey, V., Dudoit, S. & R. Irizarry, W. H., eds),. Springer New York pp. 397–420.

[Smyth *et al.*, 2003] Smyth, G. K., Yang, Y. H. & Speed, T. (2003) Statistical issues in cDNA microarray data analysis. *Methods in molecular biology (Clifton, N.J.),* **224**, 111–36.

[Snijders *et al.*, 2001] Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D. & Albertson, D. G. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature genetics,* **29** (3), 263–4.

[Solinas-Toldo *et al.*, 1997] Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Döhner, H., Cremer, T. & Lichter, P. (1997) Matrix-based comparative genomic hybridization : biochips to screen for genomic imbalances. *Genes, chromosomes & cancer,* **20** (4), 399–407.

[Song *et al.*, 2011] Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B.-K., Sheffield, N. C., Gräf, S., Huss, M., Keefe, D., Liu, Z., London, D., McDaniell, R. M., Shibata, Y., Showers, K. A., Simon, J. M., Vales, T., Wang, T., Winter, D., Zhang, Z., Clarke, N. D., Birney, E., Iyer, V. R., Crawford, G. E., Lieb, J. D. & Furey, T. S. (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome research,* **21** (10), 1757–67.

[Sorlie *et al.*, 2003] Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lø nning, P. E., Brown, P. O., Bø rresen Dale, A.-L. & Botstein, D. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America,* **100** (14), 8418–23.

[Splinter *et al.*, 2004] Splinter, E., Grosveld, F. & de Laat, W. (2004) 3C technology : analyzing the spatial organization of genomic loci in vivo. *Methods in enzymology,* **375**, 493–507.

[Srivatsan *et al.*, 2008] Srivatsan, A., Han, Y., Peng, J., Tehranchi, A. K., Gibbs, R., Wang, J. D. & Chen, R. (2008) High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. *PLoS genetics,* **4** (8), e1000139.

[Stephens *et al.*, 2009] Stephens, P. J., McBride, D. J., Lin, M.-L., Varela, I., Pleasance, E. D., Simpson, J. T., Stebbings, L. A., Leroy, C., Edkins, S., Mudie, L. J., Greenman, C. D., Jia, M., Latimer, C., Teague, J. W., Lau, K. W., Burton, J., Quail, M. A., Swerdlow, H., Churcher, C., Natrajan, R., Sieuwerts, A. M., Martens, J. W. M., Silver, D. P., Langerø d, A., Russnes, H. E. G., Foekens, J. A., Reis-Filho, J. S., van 't Veer, L., Richardson, A. L., Bø rresen Dale, A.-L., Campbell, P. J., Futreal, P. A. & Stratton, M. R. (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature,* **462** (7276), 1005–10.

[Stoeckert *et al.*, 2002] Stoeckert, C. J., Causton, H. C. & Ball, C. A. (2002) Microarray databases : standards and ontologies. *Nature genetics,* **32 Suppl**, 469–73.

[Strahl & Allis, 2000] Strahl, B. D. & Allis, C. D. (2000) The language of covalent histone modifications. *Nature,* **403** (6765), 41–5.

[Subramanian *et al.*, 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. (2005) Gene set enrichment analysis : a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America,* **102** (43), 15545–50.

[Suzuki *et al.*, 2011] Suzuki, S., Ono, N., Furusawa, C., Ying, B.-W. & Yomo, T. (2011) Comparison of sequence reads obtained from three next-generation sequencing platforms. *PloS one,* **6** (5), e19534.

[Tamayo, 1999] Tamayo, P. (1999) Interpreting patterns of gene expression with self-organizing maps : Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences,* **96** (6), 2907–2912.

[Textoris *et al.*, 2010] Textoris, J., Ban, L. H., Capo, C., Raoult, D., Leone, M. & Mege, J.-L. (2010) Sex-related differences in gene expression following Coxiella burnetii infection in mice : potential role of circadian rhythm. *PloS one,* **5** (8), e12190.

[Tomaru *et al.*, 2009] Tomaru, T., Steger, D. J., Lefterova, M. I., Schupp, M. & Lazar, M. A. (2009) Adipocyte-specific expression of murine resistin is mediated by synergism between peroxisome proliferator-activated receptor gamma and CCAAT/enhancer-binding proteins. *The Journal of biological chemistry,* **284** (10), 6116–25.

[Tusher *et al.*, 2001] Tusher, V. G., Tibshirani, R. & Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America,* **98** (9), 5116–21.

[Vazquez *et al.*, 2010] Vazquez, M., Nogales-Cadenas, R., Arroyo, J., Botías, P., García, R., Carazo, J. M., Tirado, F., Pascual-Montano, A. & Carmona-Saez, P. (2010) MARQ : an online tool to mine GEO for experiments with similar or opposite gene expression signatures. *Nucleic acids research,* **38** (Web Server issue), W228–32.

[Velculescu *et al.*, 1995] Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995) Serial analysis of gene expression. *Science (New York, N.Y.),* **270** (5235), 484–7.

[Visel *et al.*, 2009*a*] Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M. & Pennacchio, L. A. (2009*a*) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature,* **457** (7231), 854–8.

[Visel *et al.*, 2009*b*] Visel, A., Rubin, E. M. & Pennacchio, L. A. (2009*b*) Genomic views of distant-acting enhancers. *Nature,* **461** (7261), 199–205.

[Wang *et al.*, 2008] Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M. G., Glass, C. K. & Kurokawa, R. (2008) Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature,* **454** (7200), 126–30.

[Wang & Simon, 2011] Wang, X. & Simon, R. (2011) Microarray-based cancer prediction using single genes. *BMC bioinformatics,* **12**, 391.

[Wasserman & Sandelin, 2004] Wasserman, W. W. & Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nature reviews. Genetics,* **5** (4), 276–87.

[Wei *et al.*, 2006] Wei, C.-L., Wu, Q., Vega, V. B., Chiu, K. P., Ng, P., Zhang, T., Shahab, A., Yong, H. C., Fu, Y., Weng, Z., Liu, J., Zhao, X. D., Chew, J.-L., Lee, Y. L., Kuznetsov, V. A., Sung, W.-K., Miller, L. D., Lim, B., Liu, E. T., Yu, Q., Ng, H.-H. & Ruan, Y. (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell,* **124** (1), 207–19.

[Werner, 2008] Werner, T. (2008) Bioinformatics applications for pathway analysis of microarray data. *Current opinion in biotechnology,* **19** (1), 50–4.

[Wilbanks & Facciotti, 2010] Wilbanks, E. G. & Facciotti, M. T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. *PloS one,* **5** (7), e11471.

[Wilhite & Barrett, 2012] Wilhite, S. E. & Barrett, T. (2012) Strategies to Explore Functional Genomics Data Sets in NCBI's GEO Database. *Methods in molecular biology (Clifton, N.J.),* **802**, 41–53.

[Wolffe & Hayes, 1999] Wolffe, A. P. & Hayes, J. J. (1999) Chromatin disruption and modification. *Nucleic acids research,* **27** (3), 711–20.

[Wu *et al.*, 2011*a*] Wu, G., Yi, N., Absher, D. & Zhi, D. (2011*a*) Statistical quantification of methylation levels by next-generation sequencing. *PloS one,* **6** (6), e21034.

[Wu *et al.*, 2011*b*] Wu, G. P. K., Chan, K. C. C. & Wong, A. K. C. (2011*b*) Unsupervised fuzzy pattern discovery in gene expression data. *BMC bioinformatics,* **12 Suppl 5**, S5.

[Yu *et al.*, 2009] Yu, Y., Tu, K., Zheng, S., Li, Y., Ding, G., Ping, J., Hao, P. & Li, Y. (2009) GEOGLE : context mining tool for the correlation between gene expression and the phenotypic distinction. *BMC bioinformatics,* **10**, 264.

[Zeller *et al.*, 2006] Zeller, K. I., Zhao, X., Lee, C. W. H., Chiu, K. P., Yao, F., Yustein, J. T., Ooi, H. S., Orlov, Y. L., Shahab, A., Yong, H. C., Fu, Y., Weng, Z., Kuznetsov, V. A., Sung, W.-K., Ruan, Y., Dang, C. V. & Wei, C.-L. (2006) Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proceedings of the National Academy of Sciences of the United States of America,* **103** (47), 17834–9.

[Zhang *et al.*, 2004] Zhang, W., Shiraishi, A., Suzuki, A., Zheng, X., Kodama, T. & Ohashi, Y. (2004) Expression and distribution of tissue transglutaminase in normal and injured rat cornea. *Current eye research,* **28** (1), 37–45.

[Zhu *et al.*, 2010] Zhu, L. J., Gazin, C., Lawson, N. D., Pagès, H., Lin, S. M., Lapointe, D. S. & Green, M. R. (2010) ChIPpeakAnno : a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC bioinformatics,* **11**, 237.

Développement d'outils et méthodes bioinformatiques pour l'étude de l'expression des gènes et de leur régulation : Application aux pathologies

Résumé : La compréhension des mécanismes qui contrôlent l'expression des gènes est un enjeu majeur pour la recherche médicale. Elle nécessite un ensemble d'approches pangénomiques telles que les puces à ADN et plus récemment le séquençage à très haut débit qui génèrent une masse toujours plus grande de données numériques à traiter. Au cours de ma thèse, j'ai développé plusieurs outils informatiques innovants pour faciliter leur exploitation. Ainsi, j'ai créé une librairie R (AgiND) qui vérifie la qualité des données de puces à ADN Agilent et permet de les normaliser. Le nombre croissant d'expériences stockées dans Gene Expression Omnibus a motivé la mise en place du projet TBrowser. Une méthode originale DBF-MCL a été créée pour extraire des signatures transcriptionnelles annotées par l'intégration de diverses sources d'information. Stockées dans une base de données, elles sont accessibles à travers une interface Java, un service web SOAP et une librairie R/Bioconductor (RTools4TB). Enfin, un pipeline d'analyse dédié au ChIP-seq a été implémenté. Tous ces outils ont servi pour l'étude de diverses maladies dans le cadre de collaborations.
Mots clés : Bioinformatique, transcriptome, puces à ADN, épigénétique, ChIP-seq, méta-analyse

---

Development of bioinformatics tools and methods for gene expression and regulation study : Application to diseases

Abstract : Understanding the mechanisms that control gene expression is a major challenge for medical research. This requires using a large set of pangenomic approaches such as those using DNA microarrays and high-throughput sequencing that generate an ever growing mass of digital data. During my thesis, I have developed several computer-based tools to facilitate their processing and analysis. I have created a R library (AgiND) that controls the quality of Agilent DNA microarray data and allows their statistical normalization. The growing number of experiences stored in Gene Expression Omnibus has motivated the development of the TBrowser project. An original method, DBF-MCL, was created to extract annotated transcriptional signatures by integrating various sources of information. Stored in a database, these signatures are accessible using a Java interface, a SOAP web service and a R/Bioconductor library (RTools4TB). Finally, a pipeline dedicated to the ChIP-seq analyses has been implemented. All these tools were used to study various diseases in collaborations.
Keywords : Bioinformatic, transcriptome, DNA microarrays, epigenetic, ChIP-seq, metaanalysis

---

TAGC Inserm UMR_S 928, parc scientifique de Luminy, 13 288 MARSEILLE CEDEX 09